R E P O R T R E S U M E S

ED 003 009 24

THE TEST OF DIAGNOSTIC SKILLS.

BY- RIMOLDI, H.J.A. AND OTHERS
LOYOLA UNIV., CHICAGO, ILL., PSYCHOMETRIC LAB.

REPORT NUMBER CRP-1015- PT. 5 PUB DATE 62
REPORT NUMBER PUB-25
EDRS PRICE MF-$0.18 HC-$2.72 68P.

DESCRIPTORS- *MEDICAL EVALUATION, *PROBLEM SOLVING, *MEDICAL STUDENTS, TASK PERFORMANCE, PERFORMANCE FACTORS, *CLINICAL DIAGNOSIS, PHYSICIANS, CHICAGO, ILLINOIS

THE METHODS USED BY PHYSICIANS IN THE SOLUTION OF CLINICAL PROBLEMS WERE STUDIED. THIS WAS ACCOMPLISHED BY ANALYZING THE TYPE AND SEQUENCE OF QUESTIONS ASKED BY THE PHYSICIAN IN THE PROCESS OF UNDERSTANDING A PATIENT'S CONDITION. SEVEN TESTS WERE DESIGNED, TWO OF WHICH WERE THOROUGHLY EXPLORED. THESE TWO TESTS WERE ADMINISTERED SIMULTANEOUSLY TO GROUPS VARYING IN SIZE FROM APPROXIMATELY 5 TO 30 MEDICAL SCHOOL STUDENTS AND PHYSICIANS. THE SUBJECTS WERE INSTRUCTED TO LIST THE QUESTIONS ASKED DURING DIAGNOSIS. THE NATURE OF THE QUESTIONS WAS CLASSIFIED INTO THREE CATEGORIES--(1) CLINICAL INTERVIEW, (2) PHYSICAL EXAMINATION, AND (3) LABORATORY PROCEDURES. PERFORMANCE MEASURES OBTAINED SUGGEST THAT MEDICAL STUDENTS AT THE BEGINNING OF THEIR TRAINING VARY WIDELY IN THEIR APPROACH TO THE DIAGNOSTIC PROBLEM. DURING THEIR SENIOR YEAR OF TRAINING, INDIVIDUAL DIFFERENCES TEND TO BE MINIMAL. AFTER THE GRADUATED STUDENT GAINS CLINICAL EXPERIENCE, THERE APPEARS TO BE A GREATER VARIANCE IN THE NUMBER OF ITEMS SELECTED FOR A DIAGNOSIS. (WB)

ED003009

CR 1015
Part V

PSYCHOMETRIC LABORATORY

THE TEST OF DIAGNOSTIC SKILLS

U. S. DEPARTMENT OF HEALTH, EDUCATION AND WELFARE
Office of Education
This document has been reproduced exactly as received from the person or organization originating it. Points of view or opinions stated do not necessarily represent official Office of Education position or policy.

by

H. J. A. Rimoldi
John V. Haley
Hermelinda Fogliatto

LOYOLA UNIVERSITY
Chicago, Illinois

Publication No. 25

THE TEST OF DIAGNOSTIC SKILLS

by

H. J. A. Rimoldi
John V. Haley
Hermelinda Fogliatto

1962

The five year research project reported herein was sponsored by the Commonwealth Fund of New York.

Acknowledgements

The completion of this study has been possible thanks to a generous grant of the Commonwealth Fund of New York that lasted from 1956 to 1961. The principal investigators of this project express their appreciation for the opportunity afforded to them of trying novel ways of exploring medical education.

The present report summarizes research performed at the Loyola Psychometric Laboratory by one of the principal investigators and a staff that, though changing over the years, always was enthusiastic and devoted to their work.

The close cooperation with Dr. John T. Cowles, co-principal investigator of this project has been a signal factor to any success that this study may have realized.

From September, 1957, until November, 1960, Dr. Joseph R. Devane was very closely associated with this study. By looking at the references accompanying this report it is clear that his contribution to this research has been of very high value. We are greatly indebted to him for making available to us his ability and his interest in this work.

Dr. Thomas F. Grib cooperated in this study for over two years. We are very thankful for the contribution that, with his high ability, he offered towards a successful solution of some of the problems that arose in the course of this research.

We have counted on the cooperation of a bright group of graduate psychology students who were assistants in this project, among them: Dr. Stanley Cabanski, Mr. James Georgas, Mr. James Erdmann, Miss Rosemary Zacharia, Miss Isabel Reyes, Mr. Michael Partipilo and Mr. David Markert.

A decisive factor in the completion of this study has been the encouragement and interest shown by Dr. John F. Sheehan, Dean of the Stritch School of Medicine, Loyola University. He facilitated our work in every possible manner and showed a more than usual interest in an educational experiment along novel paths of inquiry. Thanks to him and to the then Secretary of the Stritch School of Medicine, we received the full cooperation of the medical faculty. To all of them we are deeply indebted.

To the Deans and Faculties of the Schools of the Health Professions, Pittsburgh, Pennsylvania, and Marquette University, Milwaukee, Wisconsin, that cooperated in this study by making available the time and experience of their staff as well as the students, who were used as subjects, we extend our sincere thanks.

In the very early stages of the development of the Test of Diagnostic Skills we counted on the sympathy and vision of an outstanding scientist, Dr. Lester R. Dragstedt, then Chairman of the Department of Surgery, University of Chicago. Thanks to his interest and scientific curiosity we moved from what was originally an idea into the field of a concrete realization.

Finally our thanks are extended to Miss Mary J. O'Connor who patiently and efficiently typed this report.

- CONTENTS -

Chapter I	- GENERAL CONSIDERATIONS ABOUT THE TEST OF DIAGNOSTIC SKILLS	1
Chapter II	- NUMBER OF QUESTIONS	7
Chapter III	- UTILITY INDEX	19
Chapter IV	- UTILITY SCORES	25
Chapter V	- AGREEMENT SCORE	36
Chapter VI	- PATTERN ANALYSIS	38
Chapter VII	- PERFORMANCE CURVES	41
Chapter VIII	- SEQUENTIAL EVALUATION OF DIAGNOSTIC PROCESSES	51
Chapter IX	- STUDIES IN PROGRESS	59
	- References	60

Introduction

The work performed for the development of the Test of Diagnostic Skills, together with other problems related to "The Development of Objective Measures of Medical Ability and Proficiency", will be described in this report. The research efforts were directed towards developing suitable instruments for appraising medical diagnostic ability, as well as appropriate scoring procedures. Many new ideas were tried, some of which are still in a developmental phase, while others have been incorporated as an integral part of the technique. New areas for research have emerged and the understanding of the diagnostic process has been partially improved and extended. Consequently it should be possible to appraise diagnostic ability more efficiently and to create those conditions that may permit to improve its training.

In the present report the results of the investigation will be synthesized, and the main lines pursued in the development of the Test of Diagnostic Skills will be described. Several publications on the different phases of the research have already appeared. A complete set of them accompanies this report.

Chapter I

GENERAL CONSIDERATIONS ABOUT THE TEST OF DIAGNOSTIC SKILLS

In the introduction of the proposal, "The Development of Objective Measurement of Medical Ability and Proficiency", May, 1956, it was stated that teaching and improving skill in medical diagnosis was one of the goals that should be pursued in medical education. The aim of this research has been to develop instruments to be used for evaluation and, if possible, training of medical students. These instruments had to be constructed to meet certain requirements of behavioral instruments. Further, they had to be realistic in terms of both medical performance and the particular framework of the educational situation.

The tests often used in order to appraise skill in medical diagnosis depend a great deal on memorizing certain items of medical information. But the problem of diagnosis also seems to be a matter of organizing the information obtained from the patient in order to reach an opinion about the case. For instance, a single check list or a collection of multiple choice questions may explore only a few of the components of the process involved in diagnosis. As a matter of fact, the diagnostic process may be viewed as a dynamic situation that changes continuously and in which the doctor-patient relationship works in a multiplicity of ways.

An exploration of the personality factors that are influential in the diagnostic situation is important. The physician's style, perception and interpretation of cues, as well as the manner and willingness of the patient to state his complaints are factors to be considered in interpreting the diagnostic situation. These aspects are not covered in the present report, though in the development of the film tests, cues concerning them were analyzed. (Cline, 1960).

In the clinical situation the diagnostic process appears to be highly structured yet very plastic. There seems to be a definite goal that the physician tries to reach, that is the diagnosis of the case. According to the nature of the case, the patient, the information and means available, and the physician, a great variety of performances can be expected. In many situations the physician proceeds by asking questions (interview, physical examination, laboratory reports) to obtain information that is brought together into a major synthesis, i.e., the diagnosis. Is it then possible, to differentiate physicians and students in terms of their ability to obtain a diagnosis and, is it possible to develop this ability?

The relationship that exists between knowledge of medical facts and diagnostic ability is not clear. Further we are not convinced that in a problem solving situation like the diagnostic process there is, or should be, a unique relationship between the diagnosis and the process followed to reach it. It is conceivable that different physicians may reach the same diagnosis using different processes or that similar processes may lead to different diagnoses. We have been more interested in the processes

than in the final result. Our study has been "oriented towards the process rather than towards the product". (Bloom and Broder, 1950).

It was thought that an approach to the study of diagnostic ability could be made by studying the sequence of questions asked by a physician to solve a clinical problem. This could give cues as to the use of the information obtained, hypotheses made, their verification, their acceptance or rejection, conditions and situations under which they are changed, etc.

The Test of Diagnostic Skills aims at exploring how a physician solves a clinical problem by analyzing the type and the sequence of questions that he asks. The basic rationale and assumptions underlying the technique have been presented in several publications. (Rimoldi, 1955, 1960).

It could be said that in the diagnostic situation the physician is actively combining and searching for information. In terms of the answers he obtains, he asks new questions and develops new hypotheses which are further examined to understand the patient's condition. Thus the diagnostic process is a dynamic situation difficult to evaluate and to control experimentally. The first and major problem was then to prepare an instrument that would not violate grossly some of the major features that seem to be important in the clinical diagnostic process and to use it as a way of appraising diagnostic ability.

Prediction always falls short of perfection unless the activity to be predicted is the same as the one used for prediction purposes. Thus the technique used in this research can not claim absolute validity. Further it is not clear that grades in medical school are satisfactory indicators of diagnostic ability, nor is there today a set of predictors that could be taken as unbiased instruments to evaluate this ability. The assumption used in this study is that if medical education develops and improves diagnostic ability and that if the Test of Diagnostic Skills estimates this ability, then changes in test performance should be related to changes in medical training.

The basic rationale of the instruments used in this study represents a departure from that used in usual testing procedures. Concepts developed in mental test theory can not always be properly applied to the Test of Diagnostic Skills. For instance, the dimension order of questions is clearly introduced in this test, and the correctness of the final answer is considered to be neither the only one nor the best index of performance. Further, the technique rather than evaluating responses stresses questions (stimuli?) freely chosen by the subjects, and is never curtailed by a multiple choice of possible final answers while the subject is being tested. Therefore a major challenge was to develop methods to evaluate these aspects. Thus it is proper to warn the users of this technique or of some of its variations that have appeared, of the difficulties that might be encountered in its use and interpretation, unless the theory and assumptions of the methods are clearly understood.

A common complaint of medical instructors is that students may have a good deal of medical information but do not know how to combine it meaningfully. To improve this would be desirable. It became apparent that the technique could be used for this purpose. For instance, by discussing with the student his performance in the test, it would be possible to make him aware of practices that are desirable or undesirable in the clinical situation. Because of lack of adequate support, we could not pursue this logical and almost compelling continuation of our research, for which we had accumulated five years of experience. We hope that this will be made possible in the future either by ourselves or by researchers aware of the possibilities, complexities and pitfalls of the technique.

The Test of Diagnostic Skills

"The subject receives, for a specific clinical case, the type of information usually available to the physician from the hospital admission chart, from the patient's complaints, and from other aspects of his clinical history. Removable cards contained in flat pockets which partially overlap are evenly arranged on a display folder. On the top edge of the numbered cards - we shall call them items - the questions that the examiner may ask are indicated. These include questions that he might wish to ask of a patient, the manipulative techniques he might wish to use, the diagnostic tests he might order, and so forth. By drawing a card and looking at the reverse side the subject gets information that is given in the form of verbal reports, laboratory analysis, X-ray films, etc. For instance, for a question like: Have you been feverish? the answer might be: Yes, yesterday afternoon had a temperature of 38°C., and for a question like: Chest X-rays, the answer might be: Both lung fields normal. Marked calcification in the arch of the aorta, and so forth.'

"Items referring to similar procedures utilized by the doctor can be grouped together, for instance those related to laboratory procedures, auscultation, endoscopy, and so forth.

"The experimenter or the subject writes the number of each item as soon as it is chosen." (Rimoldi, 1955).

The test can be used in group or individual testing. A variety of medical problems can be presented in this fashion. The experimenter may vary at will the questions as well as the amount and type of information provided, thus controlling some variables that might be influential in the final outcome. The problems selected may be real clinical cases, where actual clinical histories are transcribed totally or in part in the cards, or may be artificial cases prepared according to specified conditions, for instance to meet the requirements of particular courses or series of courses. The information can be transcribed verbally or actual observations (X-rays, electrocardiograms, etc.) may be presented without interpretation. This latter procedure has not been tried in this study. The technique is easily adapted to a variety of material and the presentation can be controlled

by the test designer.

In the present research seven tests were prepared and analyzed. Only two of them, Test 2 and 4, were thoroughly explored, and the results obtained with these will be described in the present report.

Some physicians suggested that, for evaluation purposes, made up cases may be better than real clinical cases as obtained from clinical histories. Following expert advice, Test 5 was constructed. The case presented is free of the type of information that patients often verbalize in obscure ways. A large amount of laboratory data was introduced and the test was more loaded with "factual" knowledge than what is usually found in clinical histories. In administering this test numbered folded slips of paper were used instead of cards. These slips of paper were attached to a cardboard base. Tearing them off from the base and unfolding them the subject was able to secure the information requested. The sequence of questions was obtained by requesting the subject to write down on a piece of paper the number of the slips in the order of selection. (Rimoldi, Devane, and Grib, 1958a).

The results showed that this test did not differentiate as well between levels of medical education as Tests 2 or 4 and consequently its use was discontinued. It would be difficult to attribute this to presentation, since experience acquired in other areas of inquiry with the same type of technique indicates that no major differences in performance may be attributed to the presentation of the problem. The justification of the policy of interrupting the administration of Test 5 has been presented in a previous report. (Rimoldi, Devane, and Grib, 1958a).

A short form (Form A) of both tests 2 and 4 "was prepared in order to meet the objection presented by some physicians that in a typical hospital diagnostic situation the doctor is virtually provided with a complete clinical history and physical examination data." (Rimoldi, Devane, and Grib, 1958a). For this purpose the information contained in cards referring to present illness and physical examination were eliminated from the test. Instead this information was presented on a separate piece of paper at the outset of the testing. "Otherwise, Form C (the complete form) is equal to Form A in content, format" and general instructions to the subjects. The results obtained pointed clearly that format A of both Test 2 and 4 is not as sensitive an instrument as the complete form of the same test. (Rimoldi, Devane, and Grib, 1958a).

The two remaining tests were each administered to a group of forty-five senior students. The results obtained have not been worked statistically. The reasons for this are that it was preferred to concentrate on those instruments that showed more promise than to expand into a galaxy of new tests. Rather than risk the possibility of analyzing new instruments superficially it was preferred to explore a few more intensively. This was considered justifiable because of the novelty of the technique and the

large number of problems involved in its analysis.

Test 2 and Test 4 were administered simultaneously to groups varying in size from approximately five to thirty students. Most of the physicians were tested individually, and encouraged to offer comments that could help to improve the technique. These referred to presentation, workings of the items, and sundry aspects related to both the administration of the test, the content of the test, and the instructions.

Sample of Subjects

The subjects used in this study were selected from five medical schools. Four of the schools are located in the Mid-West, three of them in the Chicago area.

The students tested were either in their junior or senior year. Table I presents the number of subjects according to school and academic level.

All of the students were tested towards the end of the school year. A large proportion of juniors of school E were tested again with Tests 2C and 4C at the end of their senior year in a longitudinal study. (Haley, 1960).

For the students of schools E and G a complete record of medical school grades and admission data was obtained.

In Chapters II through VIII the interpretation of the different scoring methods used to interpret performance in the Test of Diagnostic Skills will be presented.

Table I

Number of Subjects Who Took Each Test
in the Different Medical Schools

School	Class	Test 2C	Test 4C	Test 5	Test 2A	Test 4A
D	Juniors					
	Seniors					
	Physicians	16	16			
E	Juniors	50	45		16	16
	Seniors	24	24		22	22
	Physicians	9	8			
F	Juniors	16	15			
	Seniors	24	22			
	Physicians	6	6			
G	Juniors					
	Seniors	40	40			
	Physicians	10	10			
H	Juniors	23	29	30		
	Seniors	39	59	56		
	Physicians					
Total	Juniors	89	89			
	Seniors	127	145			
	Physicians	41	40			

Chapter II

NUMBER OF QUESTIONS

The subjects were instructed to list the questions asked in their corresponding order. From this data it is possible to compute the total number of questions asked by each subject.

The nature of the questions asked by the subjects may be classified into three categories: 1) clinical interview; 2) physical examination; and 3) laboratory procedures. For each one of these areas and for each individual, the number of questions asked was computed.

The meaning of the Number of Questions score may be difficult to interpret since it probably represents the result of a complex set of variables. The thoroughness with which a subject may explore a case may be partially inferred by the number of questions asked. Also, variation in the number of questions asked in each one of the three areas may be related to levels of training, type of training, difficulty in diagnosing the case, etc.

Tables II through XV present the means, standard deviations, and correlations for the number of questions asked. When possible, the levels of significance obtained from the statistical tests performed are indicated on the tables. When more than one comparison was made, the statistical results are presented in separate tables.

The values presented in Tables II and VII for Form A of Tests 2, 4, and 5 do not differentiate between juniors and seniors in terms of number of items selected. This holds for total number of items selected and number selected in each part. Form C of both Tests 2 and 4 does differentiate between juniors and seniors as indicated in Tables III, IV, V, VI, VIII, IX, X and XI.

Test 5 was built "ad hoc" to overcome some of the deficiencies suspected to be present in Test 2 and Test 4, which are real, complete transcriptions of actual patients. Nevertheless, Test 5 does not differentiate at any acceptable level of significance between juniors and seniors while the complete form of Test 2 and Test 4 does. This result is not surprising in view of the fact that in Form C of Tests 2 and 4 the "actual" complexity of a real case was left untouched except for those unavoidable conditions required by the technique. Test 5, being almost a textbook case, clouded any variance that may be due to the interpretation of these complexities which it is logical to assume will differ according to levels of training. Since Form A of Tests 2, 4 and 5 did not show results as promising as those found in Form C, no additional calculations were made for number of items selected. In view of the negative findings obtained using Form A it may be postulated again that for a profitable way of evaluating performance the complete clinical record should be incorporated in the test and that the subject

should be required to decide what information is necessary and when it is more useful.

For the remainder of this study Test 2, Form C, and Test 4, Form C, shall be referred to as Test 2 and Test 4. Both the first and third part of the tests differentiate at a higher level of significance between juniors and seniors than does the second part (Table IV). Experimental results indicate that juniors and seniors differ in the following order: 1) Number of items selected that refer to interview and history (Part I); 2) Number of items referring to physical data (Part II); 3) Number of items related to laboratory data (Part III).

Comparing juniors and seniors in terms of the means and variances of the number of questions asked, seniors were in general more economical and more homogeneous than juniors (Tables III and VIII). This was true in all of the schools with the exception of School H in Test 2 and 4. As the student advances from junior year to senior year, he requests less clinical interview and laboratory data. In relation to physical examination data the change is less apparent.

Physicians as a group tend to select more cards in Part I and the least number of cards in Part II. They are significantly more variable in the number of questions asked in Part I than they are in either of the other two parts in both tests. Comparing the three groups used, it can be seen that, in Part I of each test, juniors selected more cards than seniors who in turn selected more than physicians. This provides further evidence that the most striking change in the diagnostic process with clinical experience seems to be in the interview phase of the process. In the sample studied the mean number of questions asked in the interview phase of the diagnostic process appears to be a monotonically decreasing function of training and experience. For Parts II and III no discernable trend is observable. The variances followed a different trend (Table VIII). In Parts I and II and in the total of both tests, the juniors were more variable than the physicians who in turn were more variable than the seniors. In Part III of both tests, the physicians as a group were the most variable while the seniors were the least variable. The senior group then was more homogeneous than either the junior or the physician group throughout both tests.

The correlations between Test 2 and Test 4 are significantly higher for juniors and physicians than for seniors for the total and for Parts II and III. For Part I the correlation for physicians is the only one that differs significantly from the correlation for seniors (Table XIII). In general then, the junior group and the physician group resemble each other in terms of the variability and the correlations between tests more than either group resembles the seniors. On the assumption that the three experimental samples differ systematically only in training and experience, it appears that the students proceed from relative group heterogeneity and individual consistency to a stricter patterning of behavior in which the

group becomes less variable and the individual consistency less evident. After extensive clinical experience, the subjects become once again more heterogeneous and the individual's own procedures become manifest.

In other words, it appears that, as far as it can be gathered by the performance in the Test of Diagnostic Skills, medical students at the beginning of their training vary widely in their approach to the diagnostic problem. Some ask a large number of questions, while others do not. At a second stage, senior training, the students become more uniform and the seniors behave more similarly among themselves than do juniors. At this point in their training, individual differences tend to be minimal, but when the student becomes a physician and gains clinical experience, there is a greater variance in the number of items selected and individual differences again become apparent.

The fact that juniors and physicians seem alike in terms of certain statistical findings is not a picture of the real situation, but seems to reflect the effects of training upon test performance. Training will increase uniformity of procedure in as far as it would eliminate some individual differences. With further training the individual differences will tend to reappear but now they will be more closely related to the nature of the case under study, assuming the Test of Diagnostic Skills is a valid indicator of the degree of training. Thus the variability for the juniors could be interpreted as due to personality characteristics in that some will select many and some a few questions apart from the test material used. Added to this is the fact that the juniors tend to select more items as a group. This suggests that the evaluation of the information is at a rather low level. At the senior stage these individual differences tend to disappear and all the subjects act similarly. It may be inferred that at this level they are using the strict approach taught to them in medical school. Here again seniors tend to select less items. With further training and experience this uniformity is no longer evident. The physicians tend to select the least number of items, yet they vary more among themselves. Individual differences at this level could represent reasoned departures from a strict formal procedure.

The correlations between the number of questions asked in different parts of the same test do not differ significantly for the three groups as indicated by the statistical tests and by examining the columns of Table XII. This indicates that the relationship between number of cards for the three parts of the test is similar for the three groups. Correlations between number of questions asked in corresponding parts and in the totals of Test 2 and Test 4 again are all positive and significant with the exception of that for seniors in Part III (Table XIII). In Part I, this correlation is higher for physicians than for juniors or seniors. In Parts II, III, and in the total, both juniors and physicians have higher correlations between the two tests than do seniors. That is, there is a greater similarity of performance between Test 2 and Test 4 for juniors and physicians than for seniors. This is explained in part by the low

values of the standard deviations for seniors.

When the same test was administered twice to the same subjects at the junior and senior level (longitudinal study) the mean number of questions asked was smaller in the second administration than in the first (Table V). These differences were found to be significant in Part II of Test II and in Part I, Part II, and in the total of Test 4. According to these results, Test 4 appears to be more sensitive to the effects of training than Test 2, at least in terms of number of items chosen. The difference observed in Part I which refers to clinical history and interview was the greatest. In Part II, which deals with physical examination data, was next, and finally Part III, referring to laboratory techniques did not differ significantly. In general, it can be stated for Test 4 that the students tested tended to ask more questions during their junior year than they did during their senior year. This difference was most striking for questions referring to interview and clinical history. Although these same tendencies were observed in Test 2, only Part II differed significantly and that at the .05 level of confidence. The trends that were observed for standard deviations in the cross-sectional study were not evident in the longitudinal study. The variance of Part II, Test 4, was the only value that differed between juniors and seniors (Table XI). Here again juniors tended to vary more than seniors.

In the analysis of variance, the effects due to administration and to the interaction of administration by subjects were found to be significant at the .01 level of confidence. Interaction of tests by administrations was significant at the .05 level of confidence. From these results it can be concluded that the two sets of scores grouped according to administration are not from the same population of scores. In other words, the training period that took place between the first and second administration was a real source of variation for the number of questions asked.

The significant interaction effect of tests by administrations indicates that the effect of administrations on Test 2 is not an additive function of the effects of administrations on Test 4. That is, the mean change observed in Test 2 from the first to the second administration was less than the mean change observed in Test 4 from the first to the second administration. This adds further strength to the interpretation that Test 4 is more sensitive in differentiating between groups. The significant interaction of administrations by subjects indicates that the decrease in the number of items selected by each of the subjects from the first to the second administration was not constant throughout subjects. Some subjects decrease more; others decrease less. No difference was found between Test 2 and Test 4 in the first or second administration, nor was the interaction of tests by subjects significant. That is, the mean number of questions asked in Test 2 does not differ significantly from the mean number of questions asked in Test 4. Also, at a given time, each subject tended to choose the same number of items in both Test 2 and Test 4.

The correlations between questions asked in Test 2 and Test 4 in the same administration are all positive and significant (Table XIV). However, there was no significant correlations between the number of questions asked in the first and the second administration for the same test (Table XV). This indicates that in a given administration a definite relationship seems to exist between number of questions asked in Test 2 and number of questions asked in Test 4. However, for the same test there is no evident relationship between number of questions asked in the first administration and number of questions asked in the second administration.

The subjects used in this study were a highly selected group. They were very similar in age, intelligence, educational background, and possibly, even in social and environmental background. Given two tests, one after another, a subject's performance in each test would be expected to be similar if the tests themselves were similar. In this case a high correlation coefficient would be anticipated. However, if the tests were again administered after a training period and if these tests were sensitive to the type of training that took place, performance changes would be expected. It has been stated in the discussion on the analysis of variance that the rate of change from the first to the second administration was not constant throughout subjects. In view of this, and since the subjects used in this study were a highly select group, then low correlations would be expected between the first and second administrations, as shown in Table XV. This result enhances the discriminatory power of the test.

Table II

Mean Number of Cards In Each Part and The Total

	Juniors				Seniors			
	1st	2nd	3rd	Total	1st	2nd	3rd	Total
Test 2A School E	8.438	1.688	6.625	16.750	8.050	1.550	7.180	16.770
Test 4A School E	8.5000	2.125	6.188	16.812	7.409	2.500	7.318	17.227
Test 5 School H	8.900	2.330	6.930	18.170	9.732	2.286	8.625	20.643

Table III

Mean Number of Cards of Five Medical Schools In Each Part and The Total

	Schools	Juniors				Seniors				Physicians			
		I	II	III	Total	I	II	III	Total	I	II	III	Total
Test 2 Form C	E	11.200	5.500	6.720	23.420	8.040	3.420	5.170	16.620	5.889	2.667	6.333	14.889
	F	8.188	3.625	6.812	18.625	6.417	3.333	5.042	14.792	5.000	2.833	6.167	14.000
	G					9.550	5.425	7.675	22.650	6.800	4.400	7.800	19.000
	H	6.913	3.348	4.739	15.000	8.487	3.615	5.820	17.920				
	J									8.500	5.000	6.500	20.000
	Over-all Means	9.551	4.605	6.225	20.382	8.346	4.094	6.134	18.575	7.000	4.024	6.732	17.756
Test 4 Form C	E	11.733	6.533	7.067	25.333	7.500	4.792	5.500	17.792	5.625	3.000	3.375	12.000
	F	9.200	5.600	6.800	21.600	6.682	4.954	5.273	16.909	4.667	3.000	6.333	14.000
	G					8.025	5.325	6.050	19.400	6.600	6.200	8.000	20.800
	H.	9.452	5.000	5.355	19.806	9.160	5.750	5.950	20.860				
	J									9.375	5.562	6.625	21.562
	Over-all Means	10.538	5.857	6.440	22.835	8.182	5.349	5.797	19.329	7.225	4.825	6.275	18.325

Table IV

Differences Between Number of Questions Asked by the Three Groups of Five Medical Schools, in Each Part and in The Total

	School	Juniors vs Seniors				Analyses of Variance Juniors, Seniors, Physicians			
		I	II	III	Total	I	II	III	Total
Test 2 Form C	E	$p<.01$	$p<.01$	$p<.01$	$p<.01$				
	F	$p<.10$	*	$p<.01$	$p<.10$				
	G								
	H	$p<.10$	*	$p<.05$	$p<.05$				
	J								
	Over-all	$p<.10$	*	*	*	$p<.05$	*	*	*
Test 4 Form C	E	$p<.01$	$p<.05$	$p<.01$	$p<.01$				
	F	$p<.10$	*	$p<.10$	$p<.01$				
	G								
	H	*	*	*	*				
	J								
	Over-all	$p<.01$		$p<.10$	$p<.01$	$p<.01$	*	*	$p<.01$

* not significant

Table V

Mean Number of Cards Selected by Students in Each Part and in The Total, First Administration and Second Administration

N = 36		1st Admin	2nd Admin
Test 2	Part I	11.28	10.03
	Part II*	6.03	4.50
	Part III	6.64	6.36
	Total	23.94	20.89
Test 4	Part I**	12.39	8.89
	Part II*	6.61	4.81
	Part III	7.31	6.22
	Total**	26.31	19.92

First Administration greater than the Second Administration at the following levels of confidence:

* $p<.05$
** $p<.01$

Table VI

Analysis of Variance of m for 36 Medical Students
for 2 Tests and 2 Administrations

Source	Sum of Squares	df	Variance Estimate	F
Test	17.3611	1	17.3611	.6888
Administration	802.7778	1	802.7778	7.2741**
Subjects	6564.4722	35	187.5563	
Interaction: TXA	100.0000	1	100.0000	4.6384*
Interaction: TXS	882.1389	35	25.2040	1.1691
Interaction: AXS	3862.6597	35	110.3617	5.1191**
Interaction: TXAXS	754.5625	35	21.5589	
Total	12983.9722	143		

** p<.01
* p<.05

Table VII

Standard Deviation for Number of Cards
in Each Part and in The Total

	Juniors				Seniors			
	I	II	III	Total	I	II	III	Total
Test 2A School E	3.723	1.101	2.147	5.238	4.141	1.114	2.468	6.748
Test 4A School E	4.228	.857	2.067	5.331	5.340	.839	3.470	7.886
Test 5 School H	6.177	2.417	4.200	10.001	5.737	1.896	4.490	10.006

Table VIII

Standard Deviations of Number of Cards of Five Medical Schools in Each Part and The Total

	Schools	Juniors				Seniors				Physicians			
		I	II	III	Total	I	II	III	Total	I	II	III	Total
Test 2 Form C	E	5.134	3.711	2.117	8.895	4.019	1.905	1.449	5.985	5.567	2.357	2.667	9.678
	F	4.080	2.027	1.382	6.632	3.067	1.908	1.946	5.715	3.266	1.344	2.409	6.455
	G					3.486	2.673	1.506	6.167	3.544	1.800	2.857	5.254
	H	2.669	1.856	1.870	4.364	4.000	1.876	1.754	5.816				
	J									2.784	2.550	2.345	5.385
	Over-all Means	4.826	3.265	2.128	8.397	3.841	2.068	1.909	6.635	4.024	2.414	2.632	7.136
Test 4 From C	E	5.209	3.563	2.831	9.530	3.559	1.998	1.658	5.943	4.999	1.936	2.595	8.109
	F	4.020	2.304	2.166	7.309	3.349	2.035	2.434	6.495	2.495	1.528	2.982	6.028
	G					3.063	2.433	2.086	6.944	4.499	2.272	2.864	6.128
	H	4.723	2.553	2.251	8.015	4.224	2.939	2.476	7.675				
	J									3.951	2.645	2.446	7.786
	Over-all Means	5.009	3.140	2.660	8.987	3.995	2.511	2.265	7.158	4.547	2.654	3.098	8.293

Table IX

Levels of Significance for Differences Between Variances of Juniors, Seniors and Physicians Using Bartlett's Test of Homogeneity

Test	Part I	Part II	Part III	Total
2C	$p < .01$	$p < .01$	$p < .05$	$p < .01$
43	$p < .05$	$p < .05$	$p < .01$	$p < .05$

Table X

Values of t for Differences Between Variances (Correlated) of Number of Cards Selected for Each Combination of Parts for Three Groups on Two Tests

Test	Group	N	Comparison I II[1]	I III[1]	II III[2]
2	Juniors	87	4.912**	8.636**	3.932**
	Seniors	129	6.689**	8.210**	1.864
	Physicians	41	3.929**	2.916**	.573
4	Juniors	87	5.323**	7.034**	1.661
	Seniors	147	7.965**	8.070**	1.424
	Physicians	40	4.742**	2.494*	1.116

*$p < .05$.

**$p < .01$.

[1] The variance of Part I is greater than that of Part II or Part III for all groups in both tests.

[2] The variance of Part II is greater than that of Part III for Juniors and Seniors on both tests. The reverse is true for the Physicians on both tests.

Table XI

Standard Deviation of Cards Selected by Students in Each Part and in The Total, First Administration and Second Administration, Test 2 and Test 4

N = 36		1st Admin	2nd Admin
Test 2	Part I	5.1914	5.3825
	Part II	3.7304	3.0686
	Part III	2.2626	2.7904
	Total	9.3153	9.3742
Test 4	Part I	5.4226	4.3511
	Part II*	3.6000	2.4924
	Part III	2.6856	2.8491
	Total	9.9912	7.7066

First Administration greater than the Second Administration at the following level of confidence:

* $p < .05$

Table XII

Correlations Between the Number of Cards Selected in Different Parts of the Same Test for Two Student Groups and Physicians

	Test 2			Test 4		
	I II	I III	II III	I II	I III	II III
Juniors	.54**	.44**	.52**	.56**	.50**	.47**
Seniors	.58**	.40**	.47**	.66**	.37**	.37**
Physicians	.53**	.35*	.33*	.68**	.24	.51**

*$p < .05$.
**$p < .01$.

Table XIII

Correlations between the Number of Cards Selected in Corresponding Parts and the Totals of Tests 2 and 4

	$r_{I\ I}$	$r_{II\ II}$	$r_{III\ III}$	$r_{T\ T}$
Juniors (N=81)	.62**	.64**	.47**	.71**
Seniors (N=121)	.51**	.38**	.19*	.44**
Physicians (N=40)	.84**	.63**	.57**	.78**
	Ph>Jr, p<.01	Jr>Sr, p<.05	Jr>Sr, p<.05	Jr>Sr, p<.01
	Ph>Sr, p<.01	Ph>Sr, p<.05	Ph>Sr, p<.05	Ph>Sr, p<.01

*p<.05
**p<.01

Table XIV

Correlations between Number of Cards Selected in Test 2 and Test 4 for Both the First and Second Administrations

	Part I	Part II	Part III	Total
First Administration	.63**	.65**	.41*	.75**
Second Administration	.75**	.54**	.48**	.69**

** p < .01
* p < .05

Table XV

Correlations between the Number of Cards Selected in the First Administration and the Second Administration for Both Test 2 and Test 4

	Part I	Part II	Part III	Total
Test 2	.32	.15	.17	.19
Test 4	.25	.17	.07	.30

Chapter III

UTILITY INDEX

The utility index is a function of the frequency with which a question is asked by a group of subjects. It is defined as the ratio between the number of subjects who asked the question and the total number of subjects who took the test. The utility index will have different values for different questions. Also, depending on the group used for its definition, the utility index of the same question may have different values. For instance, let us assume that there is a question that is asked by a large number of physicians but seldom asked by junior students, then, in terms of the physicians, this question should have a high utility index, while in terms of the junior students, its utility index will be low. This value of the utility index can vary from 1.00 to .00.

The utility index of a question is independent of the position in which a question is asked in the sequence of questions, and thus, ignores order of selection.

Essentially the utility index represents a relative frequency of choice. It may be said that if one question is asked by many and another is asked by very few persons, then the first one may be considered to be more useful for the purposes of diagnosis. It is only in this sense that the word "utility" should be understood in this study.

The utility index has been used extensively to compare performance of different groups of people at various levels of training and to compare their performances in different tests. It defines one experimental property of the questions and it is now used to develop other types of scores. The mathematical representation of the utility index is as follows:

$$UI_{jg} = \frac{N_{jg}}{N_g} \quad (1)$$

Where UI_{jg} = utility index for card j in group g, g stands for groups, g = 1, 2, . . .g, . . .h, and j stands for card j = 1, 2, . . .j, . . .t, N_{jg} equals number of subjects in group g who selected card j, N_g equals number of subjects in group g.

If a subject's performance is analyzed using utility indexes derived from a group to which he belongs a correction should be used. In formula 1 the terms N_{jg} represents the number of subjects in the group select-item j. But a subject being evaluated in terms of the utility indexes of all the items which he selects is included in the N_{jg} term, as well as in

the N_g term. To eliminate this influence the following correction was adopted:

$$i_{jg} = \frac{N_{jg} - 1}{N_g - 1} \quad (2)$$

Where i_{jg} = corrected utility index for card j in group g. Using this correction, if a subject is the only one of a group to choose item j, the utility index of that item will be 0 and not $\frac{1}{N_g}$ as would be the case when using the uncorrected formula.
(Rimoldi, Devane, Grib, 1958a).

The utility indexes of juniors, seniors, and physicians were compared by means of X^2 Tests. In Test 2, significant differences were found between juniors and seniors for items 14, 18, 22, and 37 at the .05 level of confidence and for items 9, 28, 44, and 52 at the .01 level of confidence. In all these cases the utility indexes tended to be higher for the juniors than for the seniors. The utility indexes of the juniors differed from those of physicians for items 25 and 27 at the .05 level of confidence, and for items 5, 26, 29, 30, and 54 at the .01 level. Seniors differed from physicians for items 6, 30, 38, 45, and 52 at the .05 level and for items 54 and 56 at the .01 level.

In Test 4, the utility indexes of juniors differed from those of seniors for items 4, 9, 15, 16, 21, 22, 23, 40 and 55 at the .05 level of confidence. Items 5, 12, 14, and 27 were significant at the .01 level of confidence. Junior's utility indexes differed significantly from physician's for items 8, 12, 14, 24, 41, and 42 at the .05 level of confidence and for items 7, 10, and 15 at the .01 level. Seniors differed from physicians on items 8, 10, 24, and 42 at the .05 level and on items 7, 14, and 39 at the .01 level.

In the short form of Test 2 only two items differentiated between juniors and seniors, while in the short form of Test 4 only one item differentiated between juniors and seniors. In Test 5, although almost twice as long as Test 2 or Test 4, only five items differentiated between groups, (juniors and seniors). The long form of Test 4 again seemed to be more sensitive than the long form of Test 2 in differentiating between groups. In general, items that were significant were selected by juniors more than by seniors and by seniors more than by physicians. This is not always the case: juniors selected item 54 in Test 2 less frequently than physicians and they selected item 40 in Test 4 less frequently than seniors. In Test 2 seniors selected item 56, 52, and 45 less frequently than physicians.

In Test 4 items that differentiated between juniors and seniors also tended to differentiate between juniors and experts and between seniors and experts. In Test 2 this occurred with less items.

The medical interpretation of these differences was solicited from Dr. Frederick Selfridge from the Stritch School of Medicine. Test 2 (see Appendix) referred to an acute case that reported abdominal pain for six days duration generalized to all the right side of the body, including also the left leg, right shoulder and right hand. The patient also reported a similar pain twenty-one years ago that was diagnosed as kidney stone. The patient upon admission was a fifty-six year old butcher weighing 180 pounds, height 68.2 inches, blood pressure 125/70, pulse 96, respiration 22, and temperature 101.4 degrees Fahrenheit (38.6 degrees centigrade).

This information was given to the students beforehand, and some of the questions they asked were, in a way, irrelevant, such as item 28 which referred to accidents, item 37 referring to neck, and so forth. These did not add much information of value for diagnosing the case. Others, like item 9 referring to constipation are not usually asked of patients who show an acute abdominal syndrome. Asking about headaches (item 18) would not give directional information relating either to abdominal or thoracic organs where the pathology is presumed to reside on the basis of the information contained in the history of the present illness. In a patient of fifty-six years of age, it is usual to think in terms of degenerative or malignant processes rather than acute infectious diseases (item 27). Further, though changes in the urine, as reported by the patient, might be of value in terms of information, it is nontheless true that the type of report obtained with this question is in general highly subjective (item 22). In a patient like the one described in the test, a gastro-intestinal X-ray examination is usually precluded (item 52). The general picture seems to indicate that proctoscopy (item 56) may not be crucial to diagnose the case and though the orthopedic examination (item 44) may be recommended in terms of the distribution of the pain, it seems that, in this patient, the main symptoms refer to the gastro-intestinal tract.

Test 4 refers to a sixty-seven year old woman who for the last ten days had intermittent diarrhea and vomiting, unable to take anything by mouth. Five months before, the patient had heart trouble followed by sudden left hemiplegia with a good functional recovery in 24 hours. Upon admission the pulse was 92, blood pressure 120/60, respiration 16, temperature 97.7 Fahrenheit (36.5 degrees centigrade). The chief complaint was nausea and vomiting.

Asking about changes in appetite (item 5) in a patient with nausea and vomiting seems unnecessary. The fact that the patient had a cerebral vascular accident several months before may explain why some questions like those referring to dizziness (item 12), headaches (item 15), changes in vision (item 16) have been asked. Nevertheless the patient at the time of admission indicates a situation where the neurological components, if of interest, are not crucial. Item 14 "Have you noticed any swellings" is an ambiguous question. In a patient with nausea and diarrhea, gastroscopy (item 55) would not be expected to provide an answer for both symptoms. (Rimoldi, Devane, Grib, 1958b).

The utility indexes of the juniors of each school have been correlated with the utility indexes of the juniors of all other schools in Tests 2 and 4. The same procedure was followed for seniors and for physicians (Tables XVI, XVII, and XVIII). Also the utility indexes of the juniors and seniors of the same schools have been inter-correlated (Table XVI). The correlations obtained are all significant at the .01 level of confidence. These results seem to indicate that the three groups are closely similar in the relative frequency with which they perceive the usefulness of different items of medical information given in the tests.

If the items for the two groups under comparison have similar ranks and their relative distances from corresponding items are also similar, then the correlation coefficient as calculated by Pearson's r will be high, even if their values are different as in the case under examination. The degree of importance assigned to an item varies from physicians to seniors and juniors, but in spite of these differences the over-all approach to the problem is similar for all the groups, as indicated by the high correlations in Tables XVI, XVII, and XVIII.

In the over-all situation the long form of the tests differentiates between the groups more effectively than the short form, both at the item level and in the whole performance. Test 2 and Test 4 differentiates between groups better than Test 5, with Test 4 being the most sensitive.

The possibility exists that physicians belonging to staffs of different medical schools will show differences in their performance in the Test of Diagnostic Skills. If this is so, in order to develop valid norms, the procedure of pooling the four samples into a single group would not be permissible. The utility indexes of all the cards were calculated separately for each one of the four groups of physicians. In terms of their definition these utility indexes may be understood as an estimate of the relative importance given to each item by the persons taking the test. In order to be able to develop a meaningful scoring system, a necessary requirement is that the relative importance of the items should remain similar from group to group. In order to appraise the stability of the items from group to group the correlations of the utility indexes obtained in Test 2 and Test 4, when administered to a staff of four different medical schools were calculated. The results are given in Tables XVII and XVIII.

Simple inspection shows that for both tests the correlations are high and of about the same order of magnitude. In Table XIX, the factor matrices of the intercorrelations of Test 2 and those of Test 4 are given. It is highly interesting to note that the single factor that accounts for all the intercorrelations, both in the case of Test 2 as well as Test 4, has high loadings throughout. That is, that these four different groups of physicians perceive the relative usefulness of each item in the tests in about the same manner. Consequently, it is permissible to pool the performance of these four groups in order to develop a more stable set of utility indexes for the purposes of developing norms for scoring students.

All these findings are a strong indication that the behavior of physicians from different schools, when taken as a group is, for a specified case, highly similar.

Table XVI

Correlations of Utility Indexes (i)

Groups	Test 2	Test 4	Test 5	Test 2A	Test 4A
Juniors, Seniors, Same School					
E	.94	.91		.87	.83
F	.92	.89			
H	.94	.96	.89		
Juniors, Different Schools					
E - F	.92	.90			
E - H	.89	.93			
F - H	.88	.87			
Seniors, Different Schools					
E - F	.94	.92			
E - G	.89	.96			
E - H	.90	.93			
F - G	.91	.94			
F - H	.93	.92			
G - H	.92	.95			

Table XVII

Correlations of Utility Indexes between Physicians Schools, Test 2

Schools	J	E	F	G
J		.87	.82	.84
E	.87		.89	.83
F	.82	.88		.78
G	.84	.83	.76	

Table XVIII

Correlations of Utility Indexes between Physicians Schools, Test 4

Schools	J	E	F	G
J		.74	.73	.84
E	.74		.68	.73
F	.73	.68		.75
G	.84	.73	.75	

Table XIX

Factor Matrices of the Correlations of Table XVII and Table XVIII

	Test 2	Test 4
J	.93	.91
E	.95	.81
F	.90	.82
G	.88	.91

Chapter IV

UTILITY SCORES

The utility score is the average of the sum of the utility indexes of the cards selected by a given subject.

$$US_k = \frac{\sum_{j=1}^{m} UI_{jk}}{m_k}, \quad (3)$$

Where US_k = utility score for subject k, k equals 1, 2, . . .k, . . .N, UI_{jk} = utility index for question j, j equals 1, 2, . . .j, . . .m, and m_k = number of questions asked by subject k. Equation 3 indicates that for a subject, k, there will be a utility score US_k which is the sum of the utility indexes for all the cards he has selected divided by m_k.

Notice that the same card j will have different utility indexes according to the group used for its calculation. It follows that two individuals from different groups might conceivably select the same cards and obtain different utility scores. Therefore, utility scores as defined in Formula 3 are dependent on the group of subjects used to calculate utility indexes. This limits considerably the significance of the utility scores.

The value of the utility score varies from 0 to 1. If a subject selects questions that have not been chosen by anyone in the group used for the determination of the utility indexes, then his utility score will be 0. If all the questions he selects have a utility index of 1 then his utility score will also be 1.

The utility score can be interpreted as an estimation of the agreement of a subject with the group used for the determination of the utility indexes. The score can be computed for Part I, Part II, Part III of the tests and for the whole tests. It is clear that the utility score is a function of both the utility index of the questions asked and the number of the questions asked by the subject. In this respect it could be interpreted as indicating whether or not the subject selects questions that are considered to be very useful by the group or not. In the first case, the utility score will be high.

When utility scores are computed in terms of the performing group, corrected utility indexes should be substituted in the formula stated above, i.e.:

$$u_k = \frac{\sum_{j=1}^{m} i_{jg}}{m_k}, \quad (4)$$

Where u_k = corrected utility score for subject k, and i_{jg} is defined in (2). For more details see (Rimoldi, Devane, and Grib, 1958).

Tables XX through XXXII presented at the end of this chapter give the means, standard deviations, and correlations of utility scores for juniors, seniors and physicians in the tests administered. Since the subjects were scored in terms of various norms, a brief description of each table will be given.

Table XX presents the means of the utility scores for juniors, seniors and physicians of five medical schools, each scored in terms of their own group. These values are stated for each part and for the total of Tests 2, 4 and 5. For juniors and seniors, each individual was scored in terms of his own school. The over-all means were calculated in terms of the entire class to which the individual belonged with all schools grouped together. All of the values of the physicians, both by school and the over-all means, were calculated in terms of the entire group of physicians.

The means of the utility scores for the three groups scored in terms of physicians are presented in Tables XXI and XXII for Test 2 and Test 4 respectively. The fourth row of each table indicates the F values obtained with the analysis of variance of the corresponding columns. Table XXIII presents the means of the subjects used in the longitudinal study for the first administration and the second administration of Tests 2 and 4. The subjects were scored in terms of physicians' norms. The analyses of variance performed on these values can be found in Table XXIV.

Table XXV presents the standard deviations of the utility scores where each subject was scored in terms of his own group. These standard deviations correspond to the values used in Table XX. The standard deviations corresponding to the values in Tables XXI and XXII can be found in Tables XXVI and XXVII. Table XXVIII gives the standard deviations of the subjects in the longitudinal study. This table corresponds to the values of Table XXIII.

Correlations of utility scores calculated for each subject in terms of his own group are presented in Table XXIX. Here the different parts of the same test were correlated for each of the student groups and for the physicians. Table XXX presents the correlations between utility scores and number of questions asked by juniors and seniors in Test 2, 4, and 5. The results of the longitudinal study are presented in Tables XXXI and XXXII. These tables give the correlation between the utility scores in each test and between the utility scores of each administration respectively.

The results of Table XX indicate that the mean value for seniors was higher for each school and for all schools combined. In general the senior utility scores were also higher than those of physicians. This is a general trend in all the tests including Test 5, which was only administered to juniors and seniors of School H. This suggests that there is

more agreement among seniors than juniors as to which are the most useful items in diagnosing a case. Also there is more similarity between the values for the different schools for seniors than for juniors or physicians. The mean utility score for each part and for the total is approximately the same for all schools in the senior group. This is not the case for juniors or physicians.

In view of these results and of what has already been stated in relation to the number of items selected it would seem that seniors apply the strict approach taught in medical school and thus are in more agreement as to which items are useful. The juniors, having had a more limited training tend to agree less among themselves as to which items are useful. The lower mean utility scores for physicians could reflect a certain flexibility of approach or possible concentration on particular areas in their specialized fields.

There is no significant difference between variances for juniors, seniors and physicians when scored in terms of their own group (Table XXV). That is, the utility scores do not vary more for one group than for another.

Tables XXI, XXII, XXVI, and XXVII present the means and standard deviations of the three groups scored in terms of physicians' norms. Notice that physicians were scored in terms of corrected utility indexes, which does lower the utility scores. An independent group of physicians should be used to develop a scoring system to be applied to any subsequent group of physicians. This does not constitute a qualification of the results presented. On the contrary, the correction given in equation 4 decreases the score of the physicians. It is to be expected that the difference between the mean utility score of the physicians and those of the student groups would be even greater if a scoring key from an independent group of physicians were available.

It has been stated that the utility score can be interpreted as an index of agreement with the physicians but it can also be interpreted as an approximation to the average amount of information expected with each selection of a question. In the total test physicians tended to obtain higher scores than seniors who in turn obtained higher scores than juniors. In Part I of both tests these same significant differences were observed again with physicians getting the highest utility scores while juniors scored the lowest. This trend was also evident in Part II of Test 4. The level of significance in all these cases was at the .01 level of confidence.

Seniors were close to agreement with the physicians as to which items were useful in solving the problem. Juniors agreed less with them. This difference in agreement was most noticeable in Part I which refers to the interview phase of the test. Part III which refers to laboratory data did not differentiate between groups. There was no difference between groups in Part II of Test 2 but in Part II of Test 4 there were differences. Test 4 seems to be the most sensitive of the two tests when analyzed in terms of utility scores.

Significant differences between variances of the three groups were found only in Part III of Test II. This difference was due primarily to the seniors rather than to juniors or physicians. From this it can be assumed that, in general, the value of the utility scores did not vary more in one group than in another.

In the longitudinal study, the utility scores were computed in terms of the norms developed by the performance of a group of forty physicians. These values were higher in the second administration in each part and in the total of both tests, however, these differences were not significant in Test 2 as measured by the t test. In Test 4 the utility scores differed significantly between administrations at the .05 level of confidence in each part and at the .01 level for the total test. From this it can be concluded that the performance of the group was more in agreement with physicians' performance for the second administration.

The analysis of variance of the utility scores indicates that the mean utility score of the group did not differ significantly between tests given in the same administration. However, the differences of the mean scores were highly significant between administrations. In other words, when the group was given Test 2 and Test 4, one after another, the utility scores did not differ significantly from test to test. When these tests were administered again after a training period, although the utility scores did not differ significantly from test to test within that administration, the differences were highly significant from one administration to the next. The change in the tests' utility score from one administration to the next can be interpreted as constant, but the change in the subject's utility score from administration to administration was not. That is, in view of the significant interaction administrations by subjects, the subjects tended to get higher utility scores in the second administration than in the first, but the amount of increase was not the same for each subject.

Correlation of utility scores between tests were all significant with the exception of the first administration, Part III (Table XXXI). However, the correlations of utility scores between administrations were not significant (Table XXXII). Again a definite relationship was found to exist between utility scores of different tests within the same administration, while there is no apparent relationship between utility scores of the same test in different administrations. Here each individual performed similarly in two tests at a given period of time but after a training period performance was observed to change. According to the analysis of variance, this change in performance was not the same for each subject. Also, the group was a highly select one, and in terms of statistical theory lower values should be expected.

Table XX

Means of the Utility Scores of Five Medical Schools Scored in Terms of Their Own Groups in Each Part and the Total

	Schools	Juniors				Seniors				Physicians			
		I	II	III	Total	I	II	III	Total	I	II	III	Total
Test 2	E	.57	.60	.65	.59	.54	.51	.58	.54	.63	.59	.64	.62
	F	.66	.58	.75	.66	.62	.60	.58	.59	.68	.67	.68	.65
	G					.62	.60	.78	.66	.59	.54	.58	.56
	H	.54	.48	.59	.53	.60	.51	.65	.59				
	J									.51	.53	.67	.55
	All Schools	.59	.57	.66	.59	.61	.57	.68	.61	.58	.56	.64	.58
Test 4	E	.54	.50	.51	.54	.61	.75	.55	.60	.61	.60	.62	.57
	F	.50	.70	.65	.58	.56	.74	.53	.60	.62	.70	.58	.60
	G					.60	.63	.56	.58	.60	.59	.55	.56
	H	.63	.57	.58	.58	.61	.69	.58	.61				
	J									.48	.64	.60	.54
	All Schools	.58	.62	.57	.57	.61	.70	.57	.61	.56	.62	.59	.56
Test 5	H				.401				.447				
	All Schools				.401				.447				

Table XXI

Means of Utility Scores in Terms of Physicians for Two Student Groups (US) and Physicians (u_k) for Test 2

	Part I	Part II	Part III	Total
Juniors	.475	.545	.666	.574
Seniors	.527	.547	.671	.568
Physicians	.581	.565	.643	.583
F	8.976**	.253	2.259	4.921**

**$p < .01$.

Table XXII

Means of Utility Scores in Terms of Physicians for Two Student Groups (US) and Physicians (U_k) for Test 4

	Part I	Part II	Part III	Total
Juniors	.445	.572	.588	.504
Seniors	.525	.614	.589	.557
Physicians	.556	.625	.536	.562
F	12.545**	5.020**	.014	9.400**

**$p < .01$.

Table XXIII

Means of the Utility Scores of Students in Each Part and the Total, First Administration and Second Administration, Test 2 and Test 4

N = 36		1st Admin	2nd Admin
Test 2	Part I	.43	.47
	Part II	.50	.53
	Part III	.64	.65
	Total	.49	.55
Test 4	Part I*	.42	.49
	Part II*	.56	.63
	Part III*	.56	.59
	Total**	.48	.54

The mean US of the Second Administration is greater than the First Administration at the following levels of confidence:

* $p < .05$
** $p < .01$

Table XXIV

Analysis of Variance of US for 36 Medical Students for 2 Tests and 2 Administrations

Source	Sum of Squares	df	Variance Estimate	F
Test	.5625	1	.5625	.0079
Administration	826.5903	1	826.5903	7.7632**
Subjects	5168.9097	35	147.6831	
Interaction: TXA	70.8125	1	70.8125	1.3677
Interaction: TXS	1282.6875	35	36.6482	.7079
Interaction: AXS	3726.6597	35	106.4760	2.0566*
Interaction: TXAXS	1812.0625	35	51.7732	
Total	12746.6597	143		

** $p < .01$
* $p < .05$

Table XXV

Standard Deviations of the Utility Scores of Five Medical Schools
Scored in Terms of Their Own Group in Each Part and the Total

	Schools	Juniors				Seniors				Physicians			
		I	II	III	Total	I	II	III	Total	I	II	III	Total
Test 2	E	.096	.162	.066	.073	.115	.192	.082	.087	.171	.154	.097	.112
	F	.184	.172	.068	.121	.167	.207	.079	.109	.214	.153	.076	.103
	G					.113	.141	.068	.080	.136	.103	.062	.065
	H	.120	.180	.121	.086	.130	.143	.062	.088				
	J									.099	.134	.080	.077
	All Schools	.127	.172	.098	.096	.128	.167	.104	.096	.160	.143	.088	.088
Test 4	E	.115	.116	.06	.078	.163	.170	.094	.109	.225	.168	.109	.163
	F	.076	.135	.10	.096	.142	.162	.112	.110	.162	.088	.078	.073
	G					.176	.168	.083	.095	.133	.103	.051	.074
	H	.132	.155	.07	.092	.143	.131	.071	.092				
	J									.102	.154	.075	.082
	All Schools	.120	.155	.11	.087	.155	.161	.086	.100	.162	.144	.080	.103
Test 5	H				.076				.087				
	All Schools				.076				.087				

Table XXVI

Standard Deviations of Utility Scores for Two Student Groups Scored in Terms of Physicians (US) and Physicians (U_K) for Test 2

	N	Part I	Part II	Part III	Total
Juniors	87	.128	.171	.084	.096
Seniors	129	.135	.158	.059	.094
Physicians	41	.160	.143	.088	.088
$X^{2(1)}$		3.007	1.640	17.127**	.359

** $p < .01$

(1) Bartlett's Test of Homogeneity of Variances

Table XXVII

Standard Deviations of Utility Scores for Two Student Groups Scored in Terms of Physicians (US) and Physicians (U_K) for Test 4

	N	Part I	Part II	Part III	Total
Juniors	87	.119	.137	.086	.086
Seniors	147	.139	.130	.081	.097
Physicians	40	.162	.144	.080	.103
$X^{2(1)}$		2.549	.908	.390	2.661

(1) Bartlett's Test of Homogeneity of Variances

Table XXVIII

Standard Deviations of the Utility Scores of Students in Each Part and the Total, First Administration and Second Administration, Test 2 and Test 4

N = 36		1st Admin.	2nd Admin
Test 2	Part I	.1039	.1316
	Part II	.1501	.1240
	Part III	.0742	.0759
	Total	.0832	.0936
Test 4	Part I	.1349	.1355
	Part II	.1466	.1248
	Part III	.0721	.0799
	Total	.0956	.0899

Table XXIX

Correlations Between Utility Scores in Different Parts of the Same Test for Two Student Groups and Physicians

	Test 2			Test 4		
	I II	I III	II III	I II	I III	II III
Juniors	.36**	.35**	.51**	.13	.22*	.16
Seniors	.34**	.26**	.32**	.45**	.25**	.32**
Physicians	.30*	.23	.42**	.15	.05	-.17

* $p < .05$
** $p < .01$

Table XXX

Correlations between Utility Scores and Number of Cards Selected

	Juniors	Seniors
Test 2	-.584	-.463
Test 4	-.816	-.860
Test 5	-.725	-.745

Table XXXI

Correlations between Utility Scores in Test 2 and Test 4 for Both the First and Second Administration

N = 36	Part I	Part II	Part III	Total
First Administration	.42**	.49**	-.02	.59**
Second Administration	.39*	.48**	.33*	.42**

** $p < .01$
* $p < .05$

Table XXXII

Correlations between Utility Scores in the First Administration and the Second Administration for Both Test 2 and Test 4

N = 36	Part I	Part II	Part III	Total
Test 2	.23	-.07	.12	.03
Test 4	.23	.26	.09	.32

Chapter V

AGREEMENT SCORE

The coefficient of concordance (W) and the tau coefficient can be used to estimate agreement between series of ordered data. The tau coefficient is based on the number of inversions that occur in an ordered sequence of questions when compared to another sequence.

The coefficient of concordance indicates how several sequences agree among themselves. Therefore, if this coefficient is applied to the study of sequences followed by different groups, it is expected that its value will be higher the better the agreement among the sequences of the members of the group. (Kendall, 1955). Similarly, the higher the value of tau the greater the agreement between two sequences.

In our original approach to the characterization of the order of questions asked we decided to compute W coefficients for every test and every group (juniors, seniors, and physicians) in every school as well as for all the schools combined. The corresponding values of W are given in Table XXXIII. All these values with the exception of W for Test 5, school E are significant at the .001 level of confidence. Incidentally, comments about Test 5 in previous paragraphs show that this was a less satisfactory test.

It was postulated that a high value of W among the physicians could be a good indication that some kind of average ranking could be meaningfully used for developing some "ideal" sequence in terms of which subjects could be scored. Nevertheless, early in our work it became clear that this approach could only be accepted as a rough approximation to the interpretation and evaluation of sequences since it treats any inversion in the series the same as any other (See next chapter). For this purpose work along these lines was not pursued any further.

Inspection of Table XXXIII shows that the values of W for the shortened versions of Tests 2 and 4 (2A and 4A) tend to present values slightly lower than for the complete form of the tests. Also, Test 5, that was artificially made according to certain academic specifications, showed by far the lowest value of W.

The tau coefficient can be used to establish the relationship between the sequences of two subjects. Tau coefficients were computed for thirty-eight students who took Test 2 during both the junior and senior years. The frequency distribution of the values of the tau coefficients is given in Table XXXIV. Notice that all but six of these values are significant at least at the .05 level.

Table XXXIII

Values of W for Different Tests, Groups and Schools

School	Group	2	2A	4	4A	5
E	Juniors	.3332	.3509	.3076	.2667	
	Seniors	.4130	.3079	.4506	.2752	
	Physicians	.3845		.3145		.0424
F	Juniors	.5354		.3734		
	Seniors	.4387		.4137		
	Physicians	.5604		.5166		
H	Juniors	.4246		.4057		.2719
	Seniors	.4729		.3924		.3041
G	Juniors					
	Seniors	.4452		.3832		
	Physicians	.4971		.4899		
D	Physicians	.4459		.4188		
All Schools	Juniors	.3594		.3304		
	Seniors	.4151		.3949		
	Physicians	.4112		.3657		

Table XXXIV

Values of Tau for the Two Administrations of Test 2

Values of Tau		Frequency
.63	.679	2
.58	.629	0
.53	.579	3
.48	.529	5
.43	.479	3
.38	.429	5
.33	.379	7
.28	.329	3
.23	.279	0
.18	.229	4
.13	.179	1
.08	.129	4
.03	.079	0
-.01	.029	0
-.07	-.021	0
-.12	-.071	1

All values of .230 or more are significant at the .01 level of confidence. All values of .175 or more are significant at the .05 level of confidence.

Chapter VI

PATTERN ANALYSIS

The method of pattern analysis was devised in order to characterize patterns of responses when these responses are of a dichotomous nature. It is theoretically possible to extend the operations performed in pattern analysis to systems of more than two dimensions. For instance, one dimension representing subjects, the other, questions asked, the third, order in which the question was asked. Results have been published in which the method of pattern analysis was used for two dimensional situations. Essentially, if a group of subjects selects certain questions and not others, pattern analysis will indicate how this group agrees with a criterion group. The way of defining this is by means of an index of agreement, the value of which goes from 0 to 1 indicating respectively total disagreement and perfect agreement. Pattern analysis can be applied to a group of subjects or to a subject alone. (Rimoldi, Grib, 1960a and b).

It should be noticed that the rationale on which pattern analysis is based assumes that it is important for characterizing behavior to know both questions asked and questions not asked. Further, both questions asked and those not asked have special weight that depends essentially on their position in the pattern. The weight to be assigned to the entries in the table of responses is a function of the number of questions asked by the subject and of the frequency with which the question has been asked by the group. A question asked by a subject who asks few questions has a different weight than the same question asked by a subject who asks many questions.

The results of this method are that items are weighted differently according to the variables mentioned above rather than by assigning only weights of 1 or 0 as it is currently done. The fact that questions not asked are also used to define the pattern increases the information embodied in the final index of agreement.

Pattern analysis has been used to explore several areas of behavior besides the one of diagnostic skills. (Grib, 1961, Baldwin, 1961, Rimoldi and Grib, 1960). The basic rationale and the specific details as well as the differences that exist with other approaches used in the study of patterns have been described in the studies already mentioned. We shall not discuss this technique in any greater detail in the present report.

Work has been performed in order to determine how the system of weights alter the original rank of the response matrix. Very recently a test of significance for patterns of responses in tables of different sizes and varying number of entries has been developed. These studies have not yet been published.

The patterns of the students' responses were compared with the pattern of the physicians for both groups of students and individual students. The

physicians' pattern was established by ordering the questions in terms of their frequency of selection by the physicians. It is assumed that if a subject selects only k questions out of n possible questions, these k questions should be the same as the set of first k questions in the physicians' pattern. Any departure from this will reduce his agreement with the physicians' pattern.

In Table XXXV an arbitrary example is presented. The physicians' sequence is represented in the top row. A subject selecting three questions should choose, in terms of our definition of agreement, questions a, b and c. Subject 1 has chosen questions a, b and d, thus his agreement is less than it would be expected. Subjects 3 and 7 agree with the physicians' sequence. In the case in which the agreement is as expected, a perfect scalogram should be obtained.

Using the weights defined in previous publications (Rimoldi, Grib, 1960a and b) the weights corresponding to question d in subjects 2 and 6 are:

for subject 2: $\frac{3 \times 2}{16} = \frac{6}{16}$ and for subject 6: $\frac{2 \times 2}{16} = \frac{4}{16}$. Similarly,

for positive answers like question b in subjects 2 and 6 the weight will be respectively:

for subject 2: $\frac{4 \times 6}{33} = \frac{24}{33}$ and for subject 6: $\frac{5 \times 6}{33} = \frac{30}{33}$. That is,

the value of the entries is weighted in terms of their occurrence in relation to the question and the subject.

As described in previous publications, subjects 3 and 7 will have an index of agreement of 1.00 with the physicians' sequence.

Pattern Analysis has been applied to Test 4 only. Indexes of agreement were computed for the pattern of responses for each of 91 juniors and 143 seniors in relation to the physicians' pattern. Also, the index of agreement was determined for the group of juniors and for the group of seniors, in relation to the group of physicians. The values are respectively .7195 and .7452 indicating, as hypothesized, that the seniors' pattern agreed more with the physicians' pattern than did the juniors' pattern.

The frequency distribution of the index of agreement for each individual for both juniors and seniors approximated the normal curve ($X^2 = 13.82$ with 10 degrees of freedom and $X^2 = 15.56$ with 9 degrees of freedom respectively). The means of these distributions are .7238 and .7515 for juniors and seniors which are significantly different (P = .013 one tailed test). (Rimoldi, Haley, 1962).

The significance of each subject's performance may be determined by using the hyper-geometric distribution. The problem consists essentially

in determining the probability that when either a junior or a senior selects a certain number of questions, k or more of these will agree with the questions selected by the physicians. (Rimoldi, Haley, 1962).

Under certain specified conditions (Rimoldi, Grib, 1960a and b) pattern analysis can be applied to the direct comparison of two tables with dichotomous data, without postulating a "priori" or "ideal" sequence. The measure of significance recently developed may allow to establish the probability of a specified pattern and the differences between them.

Table XXXV

Pattern Analysis

	Physician's Sequence								
Subjects	a	b	c	d	e	f	g	Σ	Σ
1	X	X		X				3	4
2	X	X	X		X			4	3
3	X	X	X	X	X			5	2
4	X		X	X	X	X		5	2
5		X	X	X	X	X		5	2
6	X	X	X			X	X	5	2
7	X	X	X	X	X	X		6	1
Σ	6	6	6	5	5	4	1	33	
Σ	1	1	1	2	2	3	6		16

Chapter VII

PERFORMANCE CURVES

Performance curves will be defined by the accumulation of the utility indexes of the questions asked by a subject at each successive step during the process. The method allows for the placement of any individual at each stage in the process in terms of predefined maximum and minimum performance curves. It also allows for the characterization of a group in terms of the ambiguity of the questions asked and indicates how ignorance decreases at each successive step in the solution of a problem. This method has been published and applied to several problems. (Rimoldi, Devane, Haley, 1961; Tabor, 1959; Morbacher, 1960; Haley, 1960). In the present report we shall only present a summary statement about it.

In Table XXXVI under the column U_I, the utility indexes of a fictitious set of eleven questions is presented. The cumulated values obtained by adding these utility indexes from higher to lower is given in column 3 and, similarly, the values obtained by accumulating the utility indexes from lower to higher are given in colum 4. The maximum curve implies that questions are accumulated at successive steps in terms of their usefulness as expressed by the utility indexes. Similarly, the "worst" approach could be represented by cumulating the utility indexes from lower to higher. Plotting these two curves for the data in Table XXXVI, the shape indicated in figure 1 is obtained.

It is clear that if the utility indexes of all the questions have a constant value, then the area collapses into a straight line with a slope proportional to this constant value. On the other hand, if the utility indexes are different, then the questions asked can be discriminated in terms of them, and by using the procedure indicated an area will be generated between the maximum and minimum curves. The size of this area may be taken as an indication of the discriminative power of the questions in the problem. All individual performances will be enclosed within the maximum and minimum curves and at every step it is possible to indicate for each subject how far he is from them. Performances k and i as indicated in figure 1 show this clearly.

It should be clear that this characterization of performance will depend directly on the sample of subjects used to define the utility indexes. Different tests administered to the same sample of subjects should give different graphs. Thus, by keeping the sample of subjects constant and varying the tests, it is possible to obtain information about the latter. Varying the samples and keeping the tests constant information can be obtained about the subjects. This information can be expressed mathematically both in terms of areas enclosed between maximum and minimum curves or what amounts to an equivalent expression in terms of the equation of the maximum and minimum curves. In all the problems thus far examined, the equation for

the maximum curve is of the form:

$$y = c\,1 - e^{-bx} \qquad (5)$$

Where c is the y asymptote, y is sum of the utility indexes, x the number of questions asked, and b the slope. Within this formulation, the value of b indicates that the amount of information that can be gained at any stage is a constant ratio of the amount of information yet to be gained. (Rimoldi, Devane, Haley, 1961). Thus at the very beginning of the process when there is very little known, the subject should increase his information more than towards the end of the process. It is clear that if one group of subjects is administered different tests, their difficulty will be reflected by the values of b. If the same test is administered to different groups the value of b will reflect certain characteristics of the group; e.g. intelligence, level of training, etc. Smaller amounts of training should give rise to smaller values of b.

When in a group there is perfect agreement as to which items are useful in reaching a solution, then the utility indexes will be 1.00 or .00. By accumulating these utility indexes in the manner indicated above parallelograms will be obtained. The ratio between the area of the maximum and minimum curves and the area of the parallelogram indicates the degree of discrimination of the items of the test.

Assume the case of a test in which the sum of the utility indexes gives a value of 5. Assume further that the test has twelve questions. The parallelogram corresponding to this situation will be represented by a figure that accumulates a U_I of 1.00 for five questions, while for the remaining seven, the utility indexes will be 0. This has been shown in Figure 2.

If an observed performance shows that the same five questions are selected by all of the subjects and the remaining seven are selected by no subject, then the ratio between the area enclosed between maximum and minimum curves and the parallelogram will be 1.00.

If an observed performance shows that all the questions have the same U_I, then the observed performance curve corresponds to line k in Figure 2. The maximum and the minimum curves are here superimposed and the area between them will be 0. Consequently the value of the ratio will also be 0.

If an observed performance shows that the U_I of the different questions have different values, then accumulating them in decreasing order an ellipsoid as indicated in Figure 2 may be obtained. This ratio will be between 1.00 and 0, approximately 0 the less variation there is between the U_I of the cards and approximately 1.00 the greater the discrepancy between the U_I of the cards. The limiting case will be when the utility indexes are exclusively 1.00 and 0. The ratio may depend on a number of things such as difficulty of the problem, experience of the group in solving prob-

lems of a similar nature, etc.

To study performance, the utility indexes of the items were computed for the different groups. These indexes were then corrected according to formula 2 and ranked from maximum to minimum and from minimum to maximum. The graph of the cumulative utility indexes for successive choices gives the maximum and minimum performance curves. Figure 3 presents the performance curves for Test 4 and for the three groups. The parallelogram enclosing each of the performance curves represents the performance of the group if there had been perfect agreement among the group as to the expected utility of the items.

Values of b are shown in Table XXXVII. These are greater for the same group in Test 2 than in Test 4. When analysed in terms of all the scoring methods used thus far, Test 4 was found to be more difficult than Test 2. Thus it would be expected that the gain in the sum of the utility indexes for each successive question will be proportionately less for Test 4 than Test 2 as shown in Table XXXVII.

Comparing the three groups in Test 2 and in Test 4, the value of b is highest for physicians and lowest for juniors. In view of the interpretation of b as given above, the information acquired by physicians in relation to the total amount of information yet to be gained is the greatest, following next seniors and juniors in this order. In other words, physicians obtain more information at each step than seniors and seniors more than juniors.

The fact that performance in these two problems can be expressed by using the same formula may only be a coincidence and this generality awaits further consideration. However, the same type of function was also found in these same two tests in the longitudinal study. In a study of Rorschach experts diagnosing a clinical case, the same type of function was found. (Tabor, 1959). It will be of interest to know how other types of problem solving behavior may be characterized in terms of this mathematical function.

The ratios between the maximum and minimum performance curves and the corresponding parallelograms are presented in Table XXXVIII. A study of these values indicates that the ratio varies for the same test in different groups and for different tests in the same group. The ratio is always lower in Test 4 than it is in Test 2. The former increases more from juniors to physicians than the latter. Thus, for discriminating between groups, Test 4 may be considered better than Test 2. This is in agreement with the previous statements made when analyzing other scoring methods.

Figure 4 presents the performance curves that were obtained in the longitudinal study for the total of Tests 2 and 4 for both first and second administration. Similar graphs, not shown here, were also prepared for each part of each test. Again the parallelogram enclosing each of the performance curves represents the performance of the group if there had been

perfect agreement on the expected utility of the items. The ratios between observed performance curves and theoretical performance curves are presented in Table XXXIX. Again, Test 4 appears to differentiate between administrations better than Test 2. In Test 4 the greatest difference in ratios between first and second administration was observed in Part I which refers to interview and clinical history. The difference between the ratios of the two administrations of Test 2 were negligible in Parts I and III. From the inspection of Table XXXIX it may be concluded that there is more agreement for the group in the second administration than in the first, and that the most striking change is found in Part I of Test 4.

In order to get a better picture of an individual's performance in the Test of Diagnostic Skills, individual performance curves were obtained. This was done by plotting on a graph, the cumulative utility index of each subject at each successive step in the test (Figure 5). Either corrected utility indexes of the group being studied or the utility indexes of the criterion group may be cumulated. These curves indicate the number of items selected and the expected utility of these items and their sequence. At the time these curves were studied a strict method had not yet been devised which would compare individual performance curves for different tests or for different levels of training. They were introduced to see if any general statements could be made about different levels of training by their inspection.

The individual performance curves presented in Figure 5 were arrived at by plotting the cumulative utility index in terms of experts for each subject at each successive step in the test. The subjects used in this example were taken from the longitudinal study. Plateaux appeared in almost all the curves representing performance in the first administration. In the second administration, these plateaux either disappeared or became less noticeable in both tests. The items responsible for these plateaux were from Part I of the tests in every case. It seems that in the first administration the subjects' reluctance to discontinue choosing items in the interview section of the test and thus necessarily selecting items not chosen by the physicians was the primary cause for the increase in number of items selected and for the lowering of the utility scores. In the second administration the subjects were less persistant in selecting items in Part I. The plateaux disappeared or became less marked, the number of items selected decreased and the utility score increased. The examples given in the Figure were not selected because they emphasized the interpretation given above. However, they were selected from Test 4 since the plateaux were more clear-cut in this test. Thirty-one of the thirty-six subjects followed this pattern in Test 4 while only twenty-seven of the thirty-six followed it in Test 2.

Table XLVI

Data for Cumulation of Utility Indexes
in Fictitious Problem

Question Number	u_i	$\sum u_i$ max.	$\sum u_i$ min.
2	.38	.38	4.22
1	.80	1.78	3.34
8	.65	2.43	2.44
3	.50	2.93	1.79
7	.48	3.41	1.29
5	.37	3.78	.81
10	.20	3.98	.44
11	.15	4.13	.24
6	.09	4.21	.09
4	.01	4.22	.01
9	.00	4.22	.00

Table XXXVII

Absolute Values of b for three groups on
Two Tests of Medical Diagnostic Skills

	Test 2	Test 4
Junior medical students	.048	.076
Senior medical students	.059	.048
Physicians	.065	.057

Table XXXVIII

Ratios of Obtained Areas to Maximum Possible Areas

	Diagnostic Skills	
	Test 2	Test 4
Juniors	.669	.498
Seniors	.787	.618
Physicians	.820	.735

Table XXXIX

Ratio of Obtained Areas to Maximum Possible Areas

	Test 2 First Admin	Test 2 Second Admin	Test 4 First Admin	Test 4 Second Admin
Part I	.651	.650	.559	.702
Part II	.599	.675	.654	.738
Part III	.647	.642	.462	.544
Total	.649	.737	.599	.707

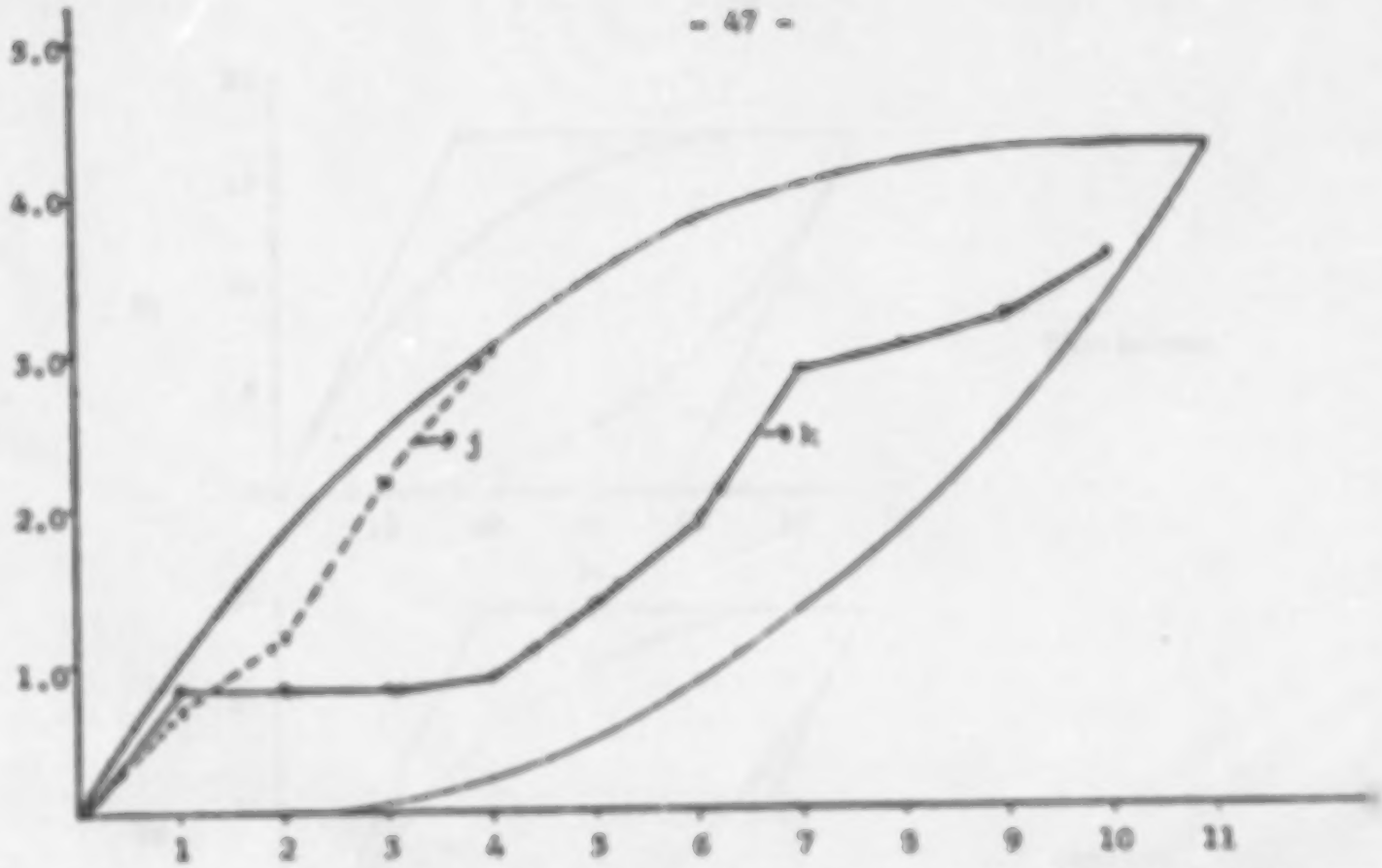


Figure 1. Maximum and Minimum Curves for Data of Table XXIVI and Performance Curves for Subjects j and k.

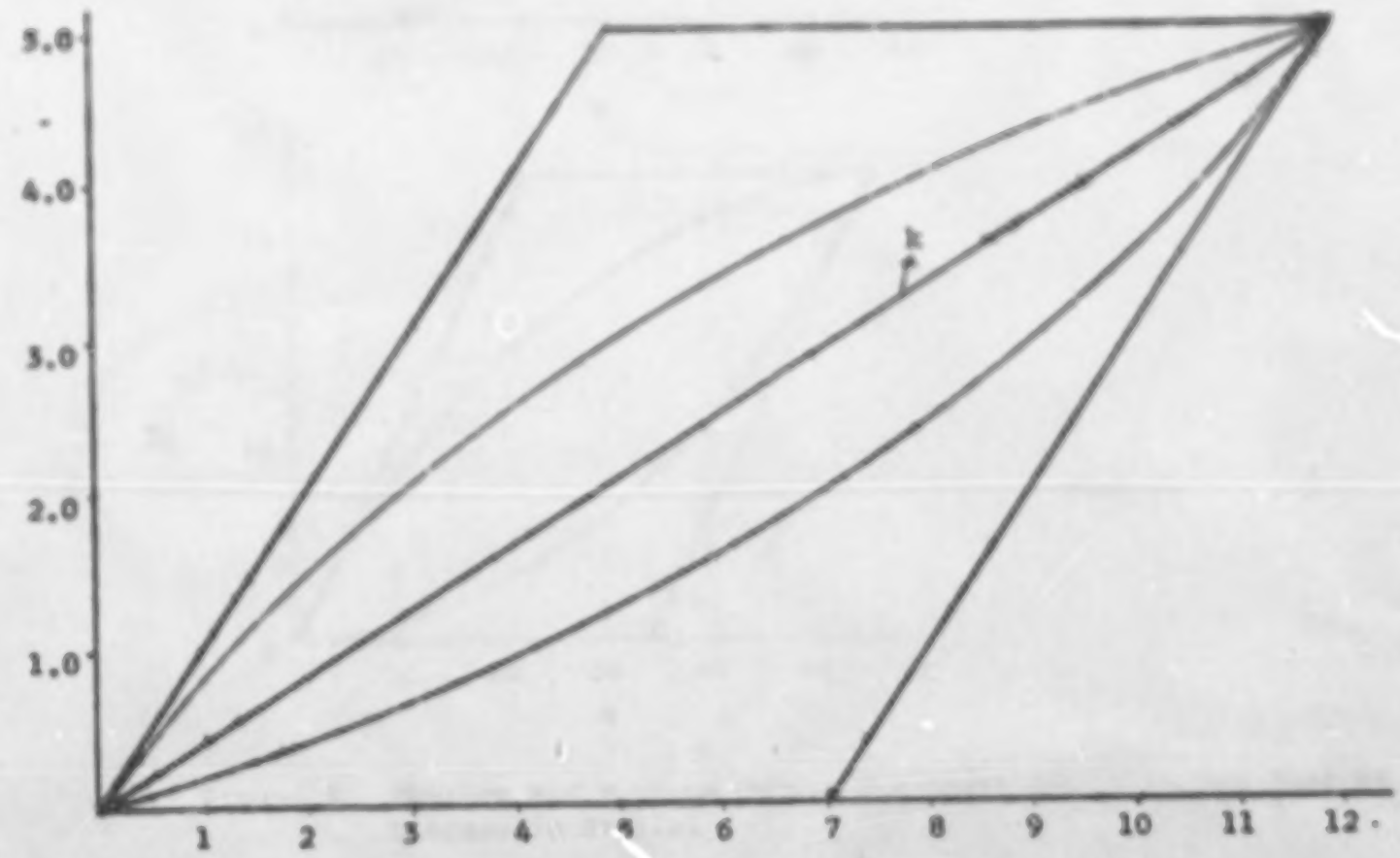


Figure 2. Maximum and Minimum Curves for Several Situations.

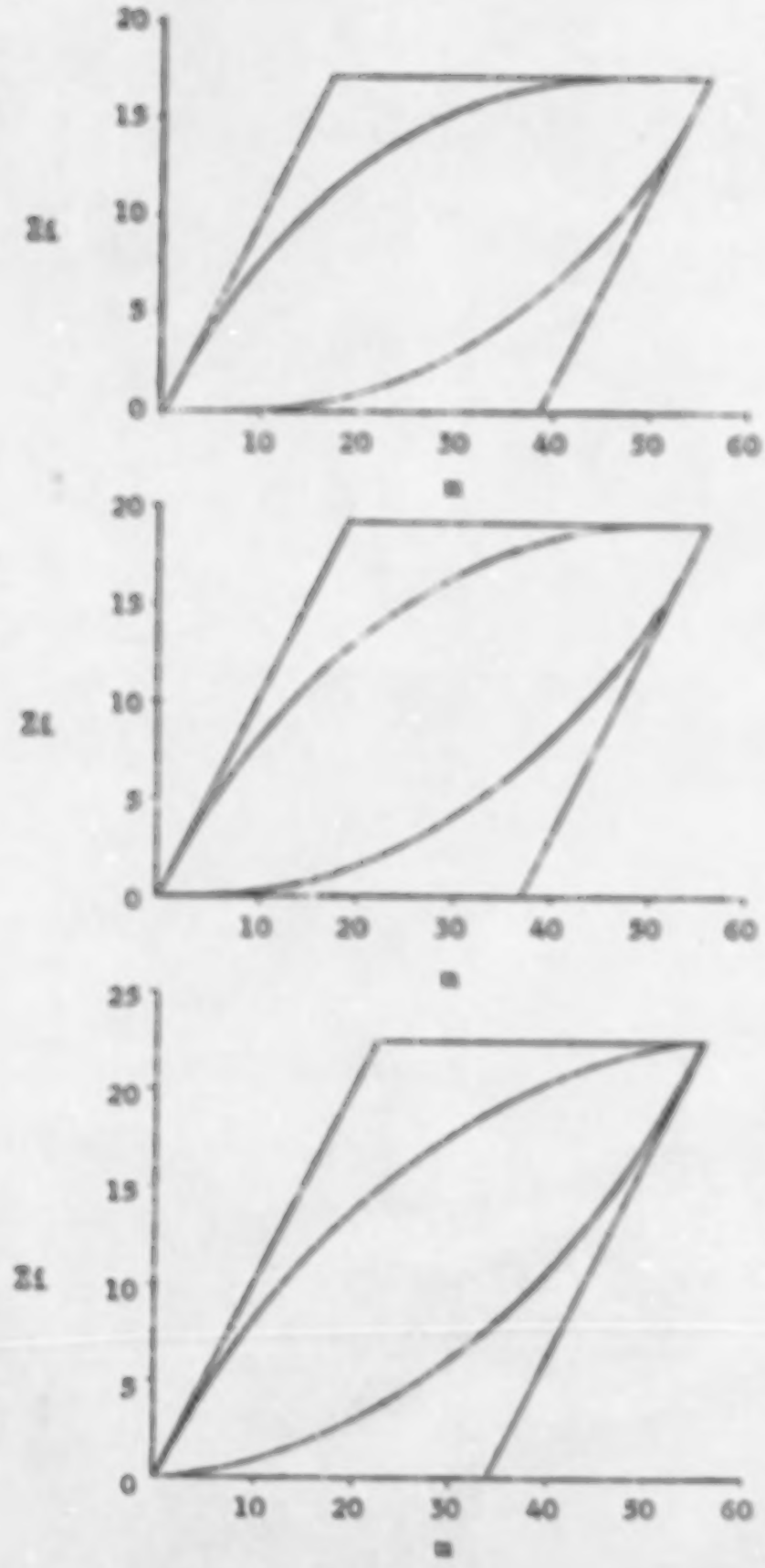
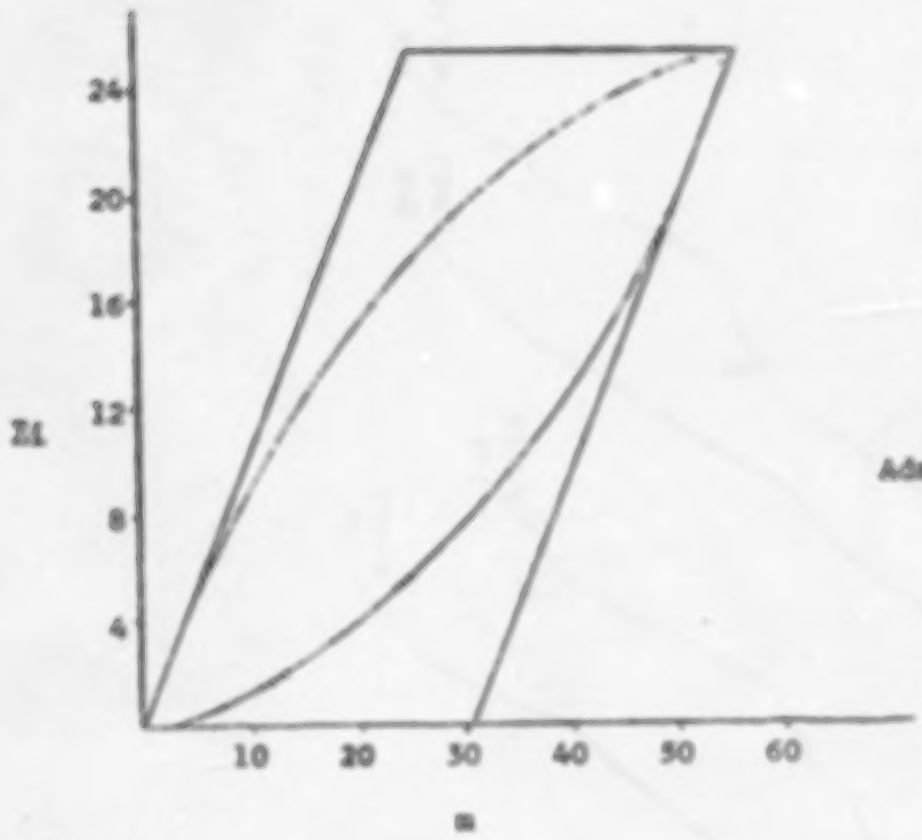
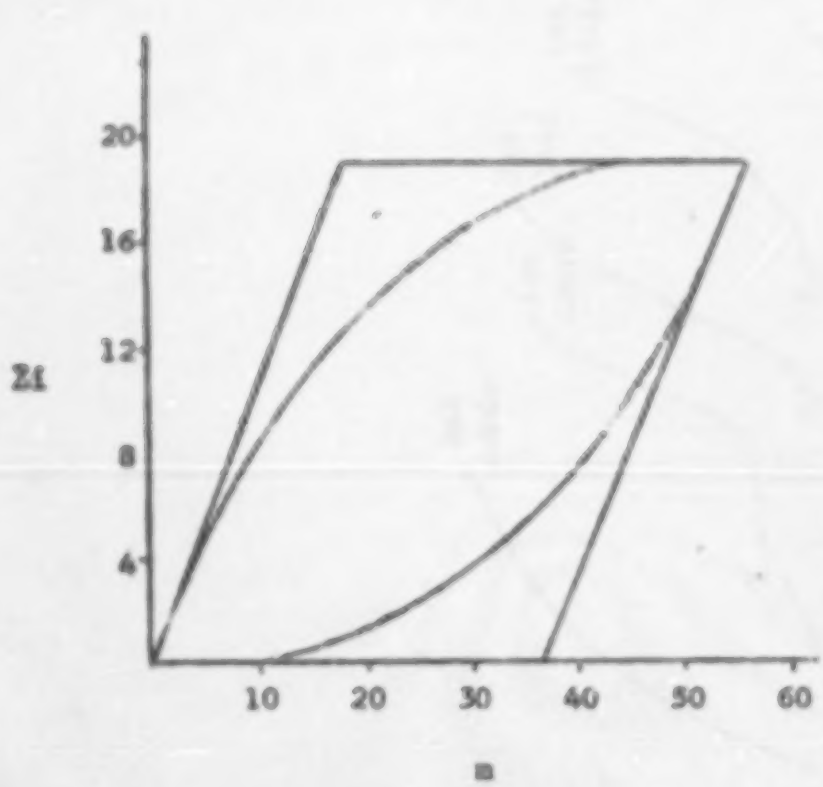


Figure 3. Maximum and Minimum Curves for three Groups on the Test of Diagnostic Skills.



First Administration



Second Administration

Figure 4. Maximum and Minimum Curves for the First and Second Administration of Test 4, Total.

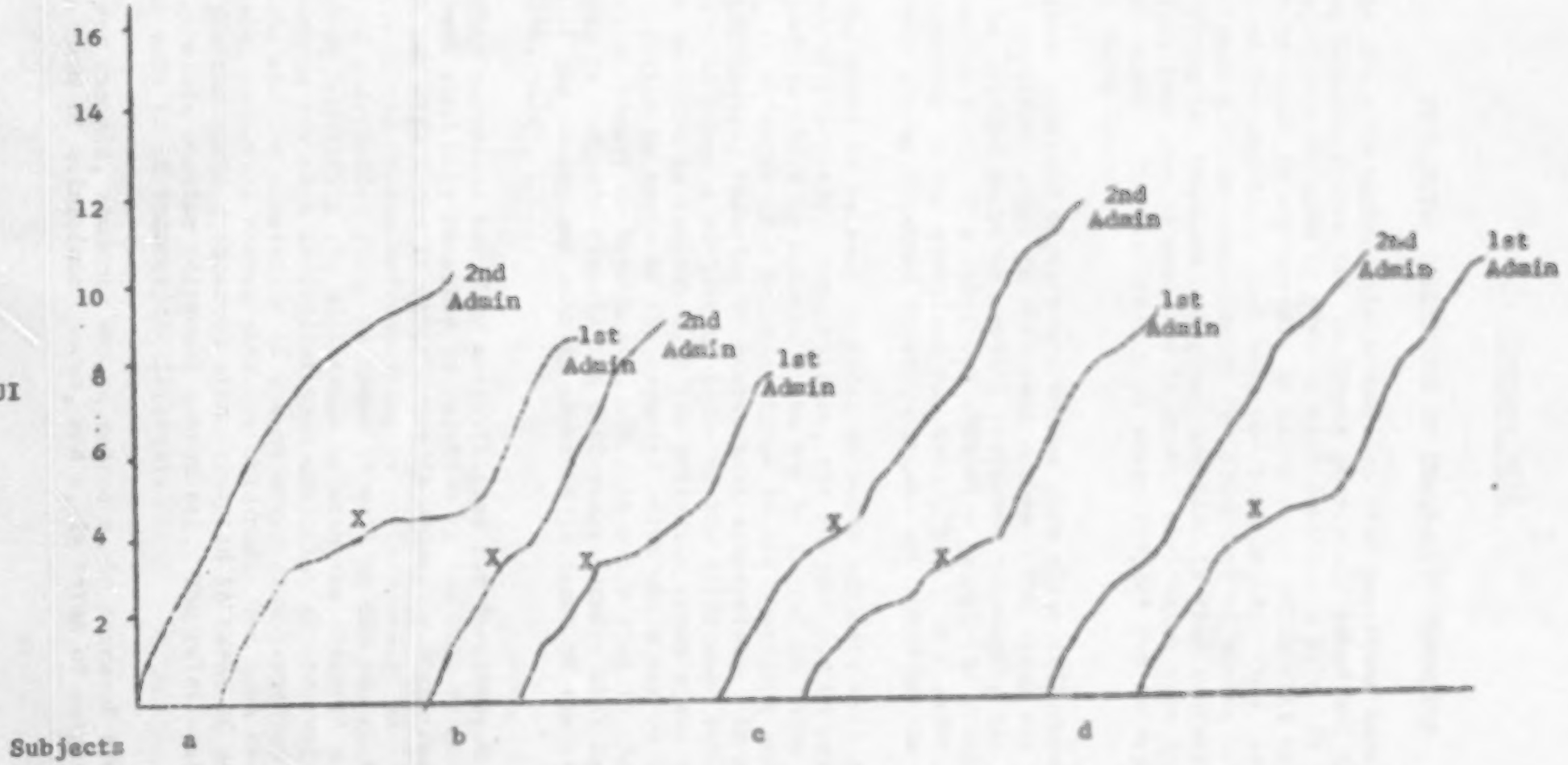


Figure 5. Individual Performance Curves for Two Administrations of the Test of Diagnostic Skills, Test 4 for Four Subjects.

Chapter VIII

SEQUENTIAL EVALUATION OF DIAGNOSTIC PROCESSES

In order to satisfy the assumption that questions have different meaning depending upon the questions previously asked and those that will follow, it was decided to give to each question a value in terms of the moment in which it was asked. An extra order which will be called 0 order refers to the questions that have not been asked. Thus, each question can be assigned a value depending on its frequency of choice at a particular time during the sequence. A given question is then characterized by a set of values that depend upon the frequency with which each question was selected in each order. The performance of every subject can be represented by cumulating these values.

These observed performance curves show clear differences for the same set of questions asked in different orders. The total sum of these values would be assumed to be an overall statement concerning the subject's performance in terms of a specified criterion group. By cumulating the values corresponding to the questions not asked, i.e., in 0 order, then a similar statement can be obtained about questions not asked by the subjects.

The norms to be used in order to score subjects will depend upon the purpose of the study. For instance, the subject may be scored in terms of the group to which he belongs or he may be scored in terms of a criterion group, or in terms of a value assigned to the questions according to specified hypotheses, theories or statistical assumptions. In all cases it is possible to score a subject in terms of the difference between the value of the question as indicated in the criterion group minus the value of the same question in terms of the expected value which may be obtained from a particular theory or hypothesis, etc. At every step in the process it is possible to indicate clearly the difference between what is expected in terms of the theory and what is observed in terms of the criterion group. (Rimoldi, Haley, 1962).

This technique has many possibilities for development some of which have been used only recently in relation to the Test of Diagnostic Skills, though our experience in other areas is wider. For instance, assume that the sum of the values corresponding to performance of an individual in terms of a criterion group is known as well as the sum of the values obtained by cumulating the difference between the observed and the expected. By plotting for each individual the cumulation of observed values on the abscissa and the cumulation of the observed minus expected values on the ordinate, polygonal convex sets are obtained. The same observed performance, when plotted against observed minus expected in terms of another hypothesis will generate another polygonal convex set. The relationship between these convex sets is of theoretical interest.

For example, students may be scored a, in terms of students' norms, b, in terms of clinicians' norms, and c, in terms of surgeons' norms. Con-

vex sets can be obtained by plotting the results obtained using students' norms and clinicians' norms and also be plotting students' against surgeons' norms. The comparison of these convex sets may help to clarify the similarities and differences in the perception of the problems by the groups. Convex sets may also be developed by plotting the results obtained using students' norms against results obtained by using both the values corresponding to students' minus clinicians' norms and students' minus surgeons' norms.

The convex sets are essentially a representation of the tactics followed by the subjects. By tactic we mean the particular sequence followed. Since there is a limited number of questions that may be asked, the number of tactics will be limited. Convex sets can also be obtained for questions not asked. It is suspected that failing to ask a question may not be the exact opposite of asking it. The comparisons of convex sets based on both questions asked and not asked seems to justify this opinion. A further interesting point is related to the fact that the same tactic may be viewed differently in terms of previous hypotheses. For example, certain tactics may show very little objective variation when explored under different hypotheses, while others may show great discrepancies. It is quite possible that tactics may be defined that would be stable under varying hypotheses, while others may not. Work on this problem began early this year.

Based on the performances of forty physicians in both Test 2 and Test 4, norms were obtained indicating the proportion of times that each question was asked by the physicians in any of the possible orders. Each question may have different values according to the place in the sequence in which it is asked by the physicians. The performance of 76 juniors and 120 seniors was scored in terms of these norms. A graphical representation of each of the subject's performance can be drawn. Accumulating the values of all the cards selected by a subject, his score can be obtained in terms of physicians' norms.

The means of these values for both Test 2 and 4 and their corresponding standard deviations are given in Table XL. There is no significant difference between the means for Test 2 and 4 of the juniors or the seniors. The correlation between the cumulative scores for both tests 2 and 4 are respectively .187 for juniors and .318 for seniors. The latter coefficient is significantly different from 0 at the .01 level of confidence (CR = 3.468). This may indicate that while in the performance of the two tests juniors show small similarity, the seniors, on the contrary, tend to perform more similarly in the two instruments.

Though in strict statistical terms it is not permissible to correlate the scores of the physicians in the two tests, because of lack of independence between the normative sample and the physicians used for the correlation, it is interesting to notice that the correlation between Test 2 and Test 4 increases to .601 in the physician group.

In the interpretation of these findings, it should be born in mind that the cumulative score is not only a function of the values of the items selected but also a function of the number of items selected. Thus in terms of a graphical interpretation of performance the same cumulative score may be the result of different performance curves.

In order to eliminate partially the effect due to number of items selected, the cumulative score for each subject was divided by the number of items selected. This was done for Test 4 in sub group of 35 students who took the test both at the junior and senior years. The mean and standard deviations are given in Table XLI.

It was hypothesized that the seniors would perform "better" than the juniors in the sense that their score would be nearer to the value of the physicians. It should be remembered that since the norms used for scoring are based on physicians' performance, the higher the score, the better the approximation to the physicians' norms in terms of frequency of choice and of order of choice in the sequence. The t test for the means in Table XLI has a value of 1.97 which for the one-tailed hypothesis is significant at the .05 level of confidence. Correlating the scores of 76 juniors, 120 seniors and 40 physicians between the two tests, the values obtained were .326, .374 and .531 respectively. All of these values were significant at the .01 level of confidence. The trend shown by these correlations moving from juniors to seniors to experts seems to indicate an increasing similarity of approach with increased medical knowledge.

Rank order correlation coefficients were computed between the cumulative scores and some grades in medical school. The results presented in Table XLII are based on 35 subjects. These subjects were tested first in their junior year and later as seniors. In both cases their performances were scored in terms of physicians' norms. Their scores as obtained in the first administration show a non-significant relationship with medical grades, but the values are higher for the second administration and for the gain from junior to senior as indicated in the last row of Table XLII. It is interesting to observe that the basic science score is the one that shows a higher correlation with performance in the test.

The results described in the previous paragraphs are an approximation to the satisfactory solution of the problem. The description of performance is not an easy matter and up to the present time, no completely satisfactory method is available for their accurate analysis.

Performance curves can be obtained for every subject using different norms. Figures 6 and 7 show the cumulative scores that represent the final values of the performance curves for Tests 2 and 4 respectively of junior and senior students scored in terms of two norms, one developed for surgeons, the other for clinicians. All of the observations below the diagonal correspond to those subjects who obtained a higher score in terms of surgeons' norms than in terms of clinicians' norms, and vice-versa for the observations

above the diagonal. For the two schools analyzed in these graphs (Schools E and H), there is a consistent trend in the students both at the junior and senior years to be on the surgeons' side in Test 2 and on the clinicians' side in Test 4. This has bearings upon the value of the two tests. Test 2 is a clear-cut case of surgical pathology while Test 4, although a surgical case, presents a complex symptomatology that may not be directly interpreted as being predominantly surgical.

Defining a subject's tactic as the sequence of questions asked by him in diagnosing a case, it is possible to compute for every test, the total number of possible tactics, which will be equal to

$$\sum_{i=1}^{n} P_i^n , \quad (6)$$

where P equals permutation, n equals the number of possible questions that may be asked and i is the number of questions asked.

Every tactic may give rise to different performance curves and different cumulative scores for every type of norm used for scoring purposes. A group of 35 students who were administered Test 4 were scored in terms of junior norms, senior norms, and physicians norms, called respectively J, S, and E. If the three norms were the same, then their differences for each question would be 0 or near 0. The performance curves for each one of these 35 students scored in the three different norms showed clear differences. By accumulating the values corresponding to all the questions asked in terms of each of the three norms, each subject will have three cumulative scores that are a function of the tactic he followed.

Figure 8 presents the following: the abscissa, corresponds to the cumulative score for Test 4 of 46 juniors scored in terms of physicians' norms (E); on the ordinate, the cumulative value of the E - J and E - S values are given. If there is a greater difference between juniors and physicians than between seniors and physicians, then the E - S values will be less than the E - J values. Each point in Figure 8 represents the cumulative score in terms of physicians' norms (abscissa) against the cumulative value of the (E - S) and (E - J) scores (ordinate). The limiting tactics have been joined by straight lines. It can be noticed that the two polygonal convex sets are different. The interpretation of these convex sets and their properties is still a matter of future research but several statements can be made. 1) The range of discrepancy is less for E - S than for E - J. 2) Some tactics show greater variation than others. Thus tactic 24 is very similar in terms of E - J and E - S, while tactic 49 shows a great discrepancy, i.e., the tactic followed by Subject 24 is more stable under the scoring methods here analyzed than the one followed by Subject 49. 3) The areas of the two convex sets are different. 4) The corners of the convex sets are not represented by the same tactics in E - J as in E - S. 5) The slope of the straight line connecting two

tactics in E - J is different than the slope connecting the same two tactics in E - S. The analogy of these findings to some concepts of game theory is unavoidable, but this will not be discussed here. (Rimoldi, 1961, Rimoldi, Haley, 1962).

Table XL

Means and Standard Deviations of Students Scored in Terms of Physicians

	Test 2		Test 4		
	Mean	σ	Mean	σ	N
Juniors	.0991	.0216	.0887	.0216	76
Seniors	.1025	.0169	.0914	.0182	120

Table XLI

The Means and Standard Deviations of the Average Cumulative Scores of Students Scored in Terms of Physicians

	Test 4	
	Mean	σ
Juniors	.0043	.0019
Seniors	.0052	.0025

Table XLII

Correlations Between Students' Cumulative Scores and Some Grades in Medical School

	Test 4			
	Three Year Av.	Senior Year Av.	Av. Basic Sciences	Four Yr. Rank
Juniors 1st Testing	-.012	.066	.079	-.001
Juniors 2nd Testing	.289	.270	.415*	.295
Juniors 2nd Testing minus 1st Testing	.233	.192	.390*	.237

* $p < .05$

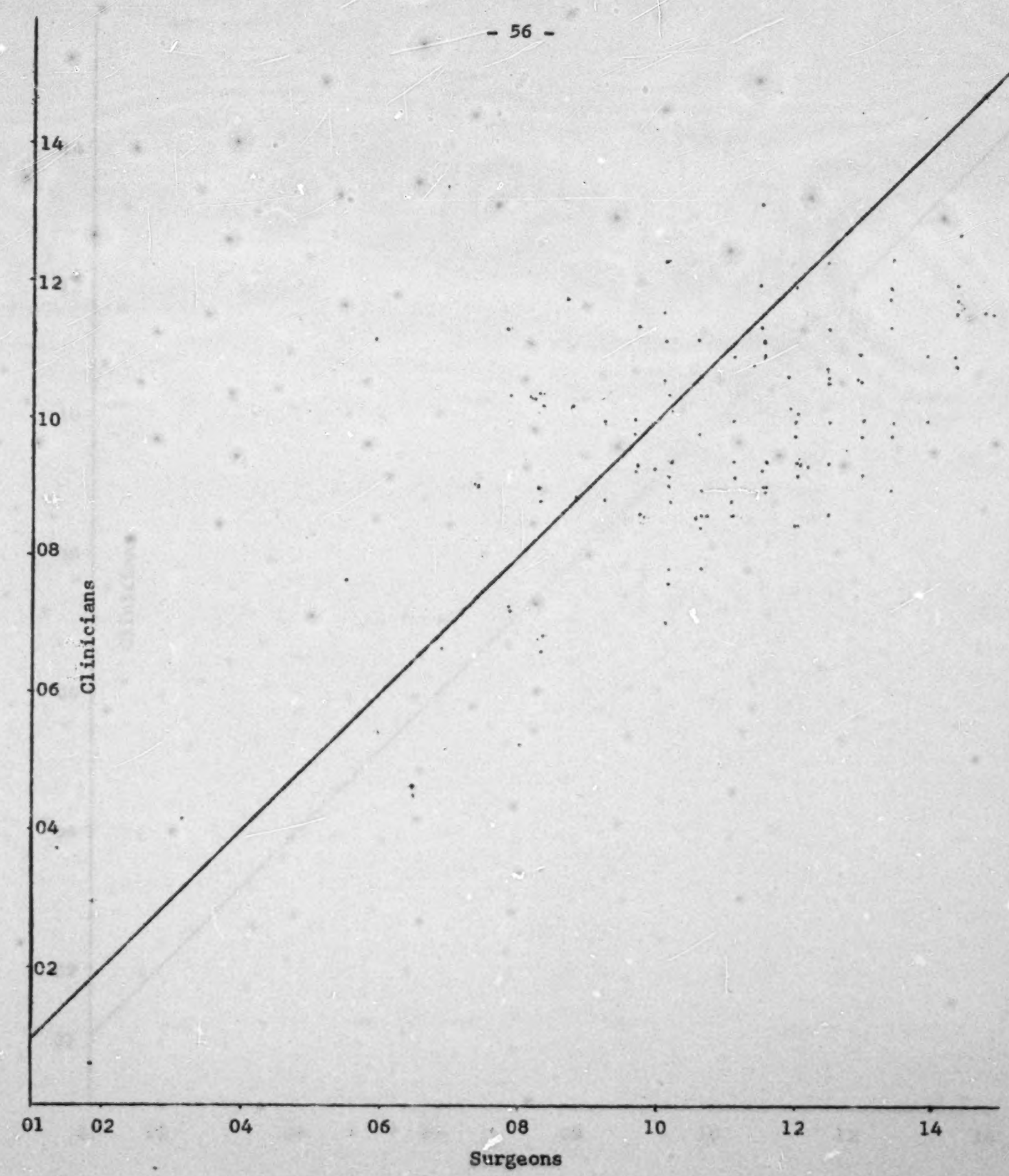


Figure 6. Cumulative Scores Representing the Final Values of the Performance Curves for Test 2 of Students Scored in Terms of Surgeons' Norms and in Terms of Clinicians' Norms.

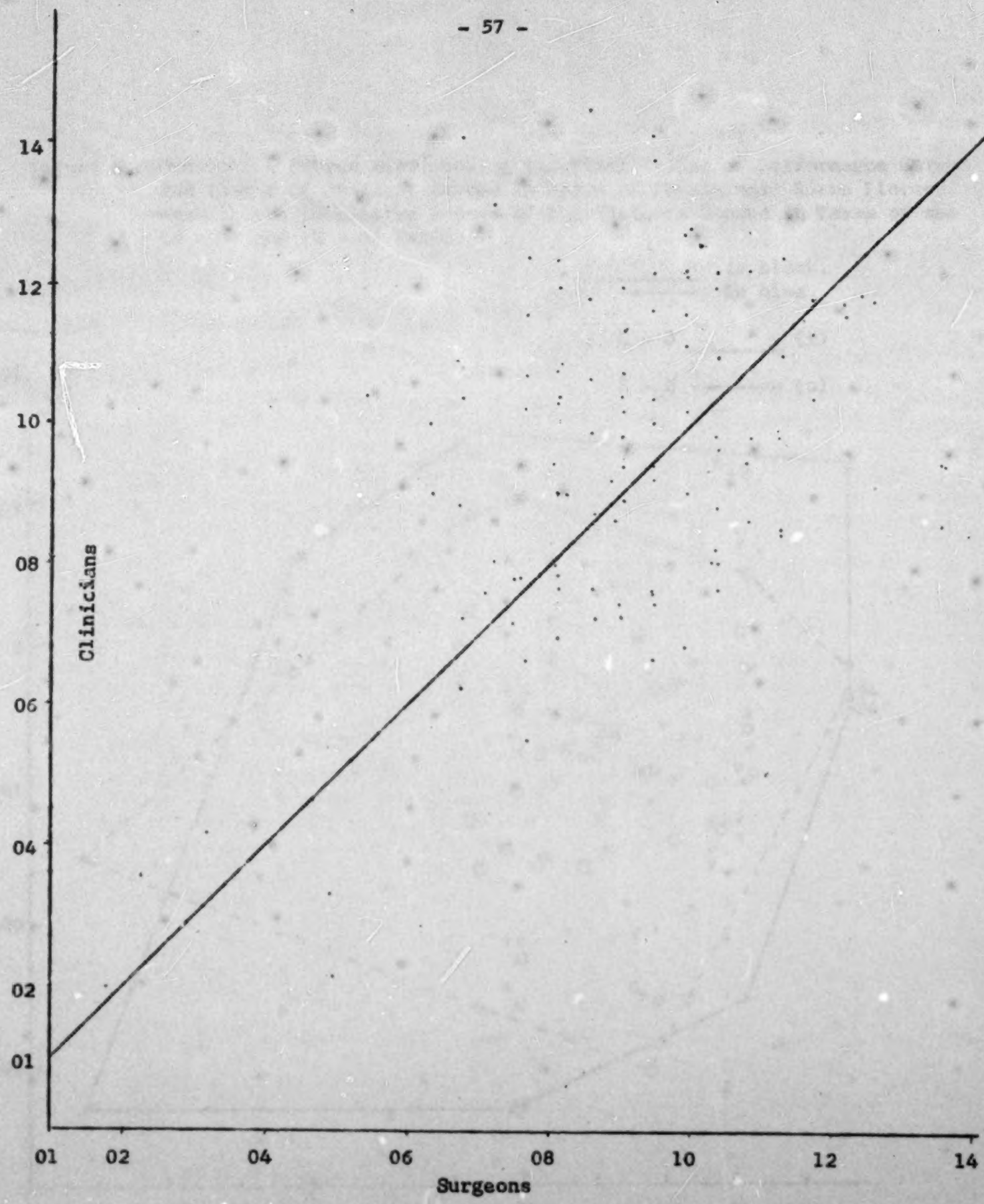
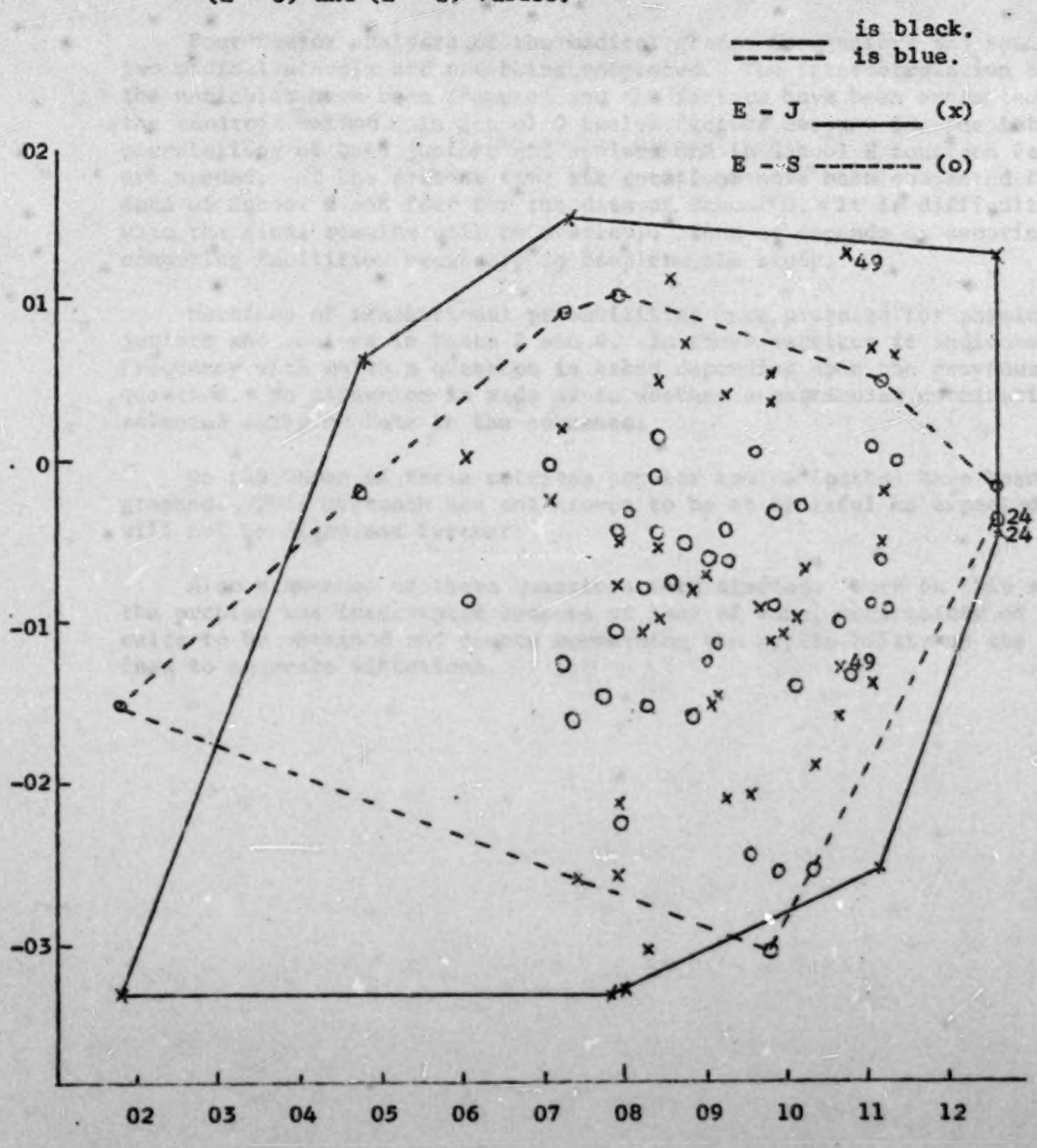


Figure 7. Cumulative Scores Representing the Final Values of Performance Curves for Test 4 of Students Scored in Terms of Surgeons' Norms and in Terms of Clinicians' Norms.

Figure 8. Cumulative Scores Representing the Final Values of Performance Curves for Test 4 of Students Scored in Terms of Physicians' Norms Plotted Against the Cumulative Scores of the Students Scored in Terms of the (E - J) and (E - S) Values.



Chapter IX

STUDIES IN PROGRESS

In this chapter a brief description of various methods of analyses that have been attempted or are in the process of completion will be presented.

Four factor analyses of the medical grades for juniors and seniors of two medical schools are now being completed. The intercorrelation between the variables have been computed and the factors have been extracted using the centroid method. In School G twelve factors account for the intercorrelations of both juniors and seniors and in School E fourteen factors are needed. At the present time six rotations have been completed for the data of School E and four for the data of School G. It is difficult to say when the final results will be available since it depends on securing the computing facilities necessary to complete the study.

Matrices of transitional probabilities were prepared for physicians, juniors and seniors in Tests 2 and 4. In these matrices is indicated the frequency with which a question is asked depending upon the previously asked question. No attention is made as to whether a particular combination was selected early or late in the sequence.

On the bases of these matrices popular routes (paths) have been graphed. This approach has not proven to be as fruitful as expected and will not be discussed further.

Also sequences of these questions were studied. Work on this aspect of the problem was interrupted because of lack of time, uncertainty of the results to be obtained and doubts concerning the applicability of the findings to concrete situations.

References

1. Bloom, B. S. and Broder, L. G., Problem Solving Processes of College Students, Chicago: The University of Chicago Press, 1950.

2. Cline, M. G., The Development of Film Tests of Skills in Clinical Observation by Medical Students. July, 1960.

3. Devane, J., Rimoldi, H. J. A., and Haley, J. V., Characteristics of Approach of Physicians to the Test of Diagnostic Skills. Loyola Psychometric Laboratory Publication No. 6, Loyola University, Chicago, Illinois., Jan. 1959.

4. Devane, J., Rimoldi, H. J. A., and Haley, J. V., A Comparison of the Performance of Two Student Groups and Physicians in the Tests of Diagnostic Skills. Loyola Psychometric Laboratory Publication No. 10, Loyola University, Chicago, Illinois., Feb. 1960.

5. Gunn, H. E., Appraisal of Personality Parameters in Terms of Processes, 1961. (Unpublished Ph.D. dissertation).

6. Haley, J. V., The Effect of Learning on Performance in the Test of Diagnostic Skills. Loyola Psychometric Laboratory Publication No. 11, Loyola University, Chicago, Illinois., Feb. 1960.

7. John, E. R. and Rimoldi, H. J. A., Sequential Observation of Complex Reasoning., American Psychologist, 10:470, 1955.

8. Kendall, M. G., Rank Correlation Methods., New York: Hafner Pub. Co., 1955.

9. Mohrbacher, J. W., An Analysis of the Interdisciplinary Evaluation of Organic Pathology in a Child Guidance Setting, 1960. (Unpublished Ph.D. dissertation).

10. Rimoldi, H. J. A., A Technique for the Study of Problem Solving. Educational and Psychological Measurement, 1955, 15, 4, 450-461.

11. Rimoldi, H. J. A., Problem Solving as a Process. Loyola Psychometric Laboratory Publication No. 5, Loyola University, Chicago, Illinois, 1958. Educational and Psychological Measurement, 20, 3, 1960.

12. Rimoldi, H. J. A., The Test of Diagnostic Skills. Loyola Psychometric Laboratory Publication No. 15, Loyola University, Chicago, Illinois, Jan. 1960. Journal of Medical Education, 36, 1, Jan. 1961.

13. Rimoldi, H. J. A., L'Étude Des Processus Psychologiques. Travail Humain, France, 1961. The Study of Psychological Processes. Loyola Psychometric Laboratory Publication No. 19, Loyola University, Chicago, Illinois, 1961.

14. Rimoldi, H. J. A., and Devane, J., Inner Organization of the Clinical Diagnostic Process as Appraised by the Test of Diagnostic Skills. Loyola Psychometric Laboratory Publication No. 4, Loyola University, Chicago, Illinois, 1958.

15. Rimoldi, H. J. A., Devane, J., and Grib, T. F., Testing Skills in Medical Diagnosis. Loyola Psychometric Laboratory Publication No. 2, Loyola University, Chicago, Illinois, 1958a.

16. Rimoldi, H. J. A., Devane, J., and Grib, T. F., Supplementary Report on Testing Skills in Medical Diagnosis. Loyola Psychometric Laboratory Publication No. 3, Loyola University, Chicago, Illinois, 1958b.

17. Rimoldi, H. J. A., Devane, J., and Haley, J., Characterization of Processes. Loyola Psychometric Laboratory Publication No. 8, Loyola University, Chicago, Illinois, 1959. Educational and Psychological Measurement, 1961.

18. Rimoldi, H. J. A., and Grib, T. F., Pattern Analysis. Loyola Psychometric Laboratory Publication No. 7, Loyola University, Chicago, Illinois, 1958. British Journal of Statistical Psychology, Nov. 1960.

19. Rimoldi, H. J. A., and Grib, T. F., Some Properties and Applications of Pattern Analysis. Loyola Psychometric Laboratory Publication No. 14, Loyola University, Chicago, Illinois, 1960.

20. Rimoldi, H. J. A., and Haley, J., Sequential Evaluation of Problem Solving Processes. Loyola Psychometric Laboratory Publication No. 22, Loyola University, Chicago, Illinois, 1961.

21. Rimoldi, H. J. A., and Haley, J. V., Determining Significance Levels in Pattern Analysis. Loyola Psychometric Laboratory Publication No. 23. Psychological Reports, 10, 500, 1962.

22. Tabor, A., Process Analysis of Rorschach Interpretation, 1959. (Unpublished Ph.D. dissertation).